A Probabilistic Method for Tracking a Vocalist

Lorin V. Grubb

September 1998

CMU-CS-98-166

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Roger B. Dannenberg, Chair
Tom Mitchell
Jack Mostow
Shuji Hashimoto, Department of Applied Physics, Waseda University, Tokyo

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

Keywords: Computer Music, Stochastic Processes, Signal Processing, Singing, Musical Accompaniment, Statistical Modeling, Automated Accompaniment

Abstract

When a musician gives a recital or concert, the music performed generally includes accompaniment. To render a good performance, the soloist and the accompanist must know the musical score and must follow the other musician's performance. Both performing and rehearsing are limited by constraints on the time and money available for bringing musicians together. Computer systems that automatically provide musical accompaniment offer an inexpensive, readily available alternative. Effective computer accompaniment requires software that can listen to live performers and follow along in a musical score.

This work presents an implemented system and method for automatically accompanying a singer given a musical score. Specifically, I offer a method for robust, real-time detection of a singer's score position and tempo. Robust score following requires combining information obtained both from analyzing a complex signal (the singer's performance) and from processing symbolic notation (the score). Unfortunately, the mapping from the available information to score position does not define a function. Consequently, this work investigated a statistical characterization of a singer's score position and a model that combines the available musical information to produce a probabilistic position estimate. By making careful assumptions and estimating statistics from a set of actual vocal performances, a useful approximation of this model can be implemented in software and executed in real time during a musical performance.

As part of this project, a metric was defined for evaluating the system's ability to follow a singer. This metric was used to assess the system's ability to track vocal performances. The presented evaluation includes a characterization of how tracking ability can be improved by using several different measurements from the sound signal rather than only one type of measurement. Examined measurements of the sound signal include fundamental pitch, spectral features dependent upon the score's phonetic content, and amplitude changes correlated with the start of a musical note. The evaluation results demonstrate how incorporating multiple measurements of the same signal can improve the accuracy of performance tracking, for individual performances as well as on average. Overall improvement of the performance tracking system through incremental specification, development, and evaluation is facilitated by the formal statistical approach to the problem.

Acknowledgments

First and foremost, I would like to thank my advisor, Roger Dannenberg, for both his assistance and his support, as well as for his great patience throughout the life span of this project. Thanks are also due to the members of my thesis committee—Shuji Hashimoto, Jack Mostow and Tom Mitchell—for their numerous insights and helpful comments as well as their general interest in this work. I am also grateful for the many helpful discussions with other faculty and students in the Department of Computer Science including John Lafferty, Rich Stern, Lonnie Chrisman, Sven Koenig, Ravi Mosur, Eric Thayer and Belinda Thom.

I am indebted to numerous faculty and students in the Department of Music at Carnegie Mellon University. My efforts would have been far less fruitful without their support for this project and their willingness to devote generous amounts of time to participate in the conducted studies. Special thanks are due to Anne Elgar Kopta, Robert Fire and Geeta Bhatnagar.

Finally, I wish to thank my parents for their encouragement and support during my extensive time as a student in the areas of both science and music.

Contents

Figures

Tables

Chapter 1

The Problem of Vocal Performance Tracking

1.1 Musical Performance and Computers

This work explores one important aspect of applying computers to musical performance. Performance is an important part of both music and music education. Large numbers of people attend concerts, go to live performances in bars and clubs, listen to music broadcast over radio and television (perhaps soon the internet?), and purchase recordings. Students of music perform regularly in recitals and concerts, and in addition may simply "make music" in small, private groups purely for enjoyment. All of these musical experiences, in one form or another, involve performance.

In the tradition of Western classical music, a performance usually involves the rendering of a previously constructed and relatively well-specified work, or *composition*. The entirety of a musical composition is notated fairly explicitly in a *musical score*, or simply a *score*. An excerpt from a musical score is shown in Figure 1-1. The score contains several *staves*, each staff containing notes and other information that describe the performance to be rendered on a given instrument. The piece in Figure 1-1 is written for voice and piano. The staves are grouped in sets of three. The upper staff in each group describes the *vocalist's part*, or the performance to be rendered by the singer. The lower two staves in combination describe the pianist's part. Location of a note head on the staff indicates pitch. The appearance of the note, both the head and the beam in combination, indicate relative duration of the note. For instance, a note with a solid head and no beam (a *quarter note*) ideally spans twice the time spanned by a note with a solid head and one beam (an *eighth note* or *quaver*). Notes that are vertically aligned across the parts within a staff group are intended, in a literal rendering of the score, to be performed simultaneously. For pieces such as the one in this example, the part to be sung is often referred to as the *soloist's part*. The part to be performed by the pianist is referred to as the *accompaniment*.

During an actual performance, the soloist (the singer) and the accompanist (in this example, the pianist) must work together. They must listen to and respond to one another in order to generate a performance that is musical and aesthetically acceptable. The performers must strive to synchronize their

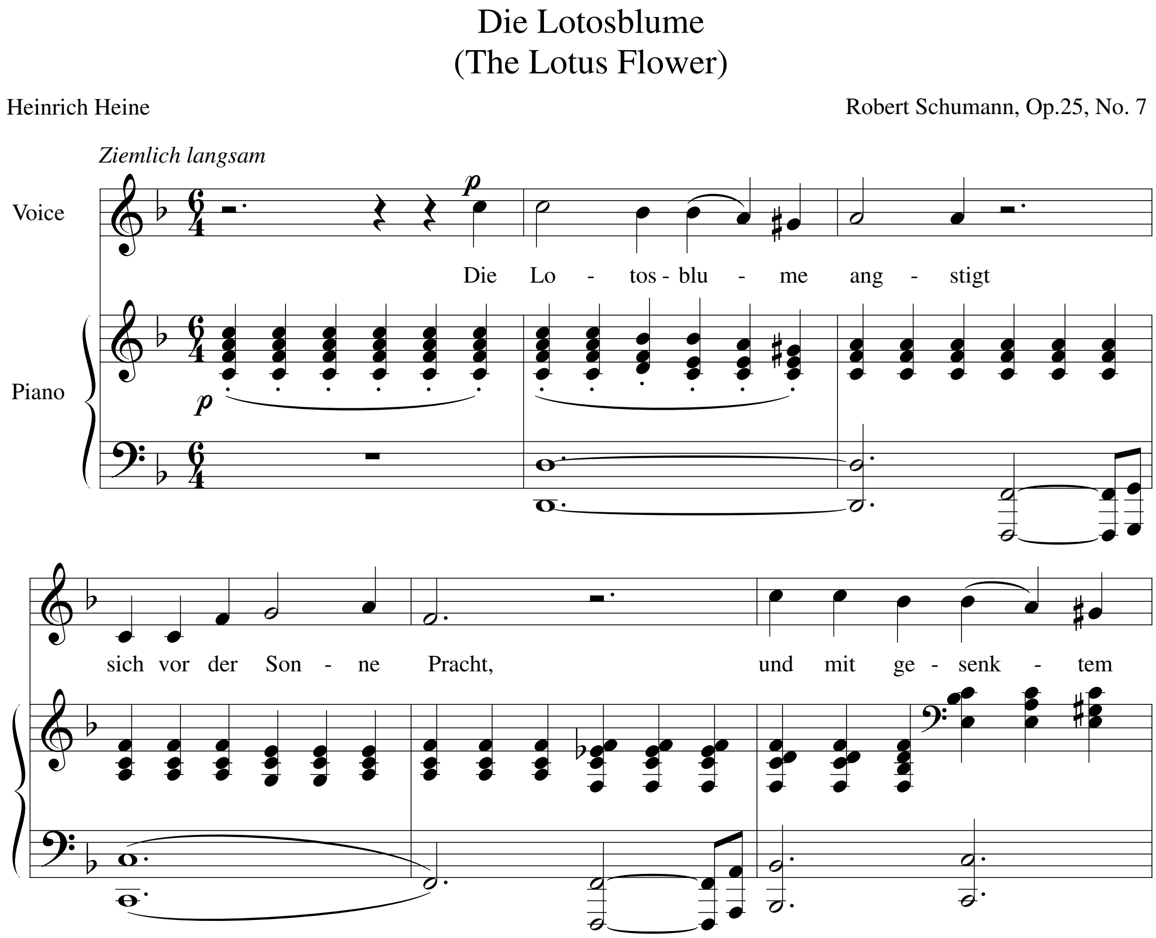


Figure 1-1. Excerpt from a musical score of a piece for voice and piano.

individual performances at the appropriate points in the score. They must agree upon the behaviors and the roles of each performer at different points in the score, and they must avoid violating these expectations to any serious degree. Needless to say, the details of these behaviors and expectations are numerous and interact in complex ways. Competent musicians work diligently to acquire these details as part of their musical training, and proper execution of a musical performance depends heavily upon the knowledge and experience of the individual performers.

Computer systems capable of recording, processing, and generating sound have been available for several decades. These systems have found numerous uses within the field of music, and their evolution has often been motivated by their application to music. One goal of applying computers in music is to develop systems capable of providing quality accompaniment. These systems are often referred to as *automated accompaniment systems*. A simple block diagram of an automated accompaniment system is presented in Figure 1-2. Ideally, accompaniment systems should accept a symbolic score that contains all parts in the piece, monitor a performance by a live soloist, and perform the other parts of the score in real time and in a musical manner. This latter requirement demands an ability to synchronize with the live performer and to adjust the performance in response to the soloist. For instance, as the soloist increases the tempo (plays faster), the accompaniment system should do likewise where appropriate.

Use and construction of automated accompaniment systems is motivated by the limitations of other means for generating musical accompaniment. A performer desiring accompaniment has essentially two choices. First, other live musicians are obviously capable of accompanying the performer. However, use of live accompaniment is limited by both constraints on scheduling the live musicians and the cost of hiring them. These limitations apply to accompaniment by a single pianist as well as by larger groups such as chamber ensembles and orchestras. Second, a performer may choose to rehearse or perform with a recording rather than other live performers. While recordings are both affordable and permit unlimited use whenever desired, they limit the musical expressiveness of the live performer. Recordings do not respond to the soloist, who will have to assume a tempo and dynamic level appropriate to the static recording. Furthermore, if the only available recording includes performance of the soloist's part, the performer may be so distracted by the competition as to find rehearsal impossible.

Consequently, automated accompaniment systems are useful to soloists. These systems are more readily accessible than live musicians and can be more responsive to the performer than static recordings. Utility of computer accompaniment is further confirmed by the existence of at least one commercially available system. Produced by Coda Music Technology, Inc., the Vivace Intelligent Accompanist™ is an application that runs on personal computers and provides real-time playback of a musical accompaniment. The system was designed primarily for educational uses, and is purchased primarily for high school

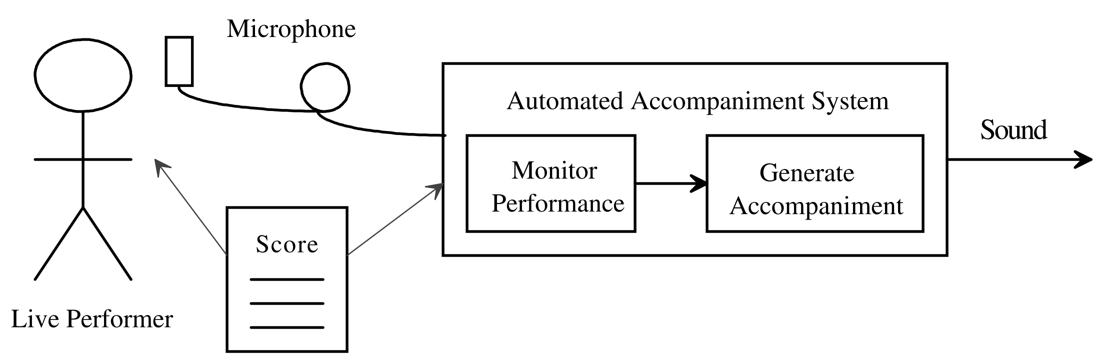


Figure 1-2. Simple block diagram of an automated accompaniment system.

students and individuals studying performance on a musical instrument. Sales of the system have been reasonable, exceeding 2000 units. In addition, certain competitions sponsored by state music organizations have enthusiastically permitted use of the system in lieu of a live accompanist.

The work undertaken in this study is motivated by the desire to construct software systems for automated musical accompaniment. The primary interest is in understanding how to construct robust systems capable of accompanying vocal performers. The goal is made even more ambitious by targeting accompaniment systems that are reasonably general—systems that, with little or no effort beyond input of a new score, can accompany different singers of similar skill performing many different pieces exhibiting multiple styles and genres. Beyond musical aesthetics, issues of both generality and reliability are of primary concern. The next section provides a more complete and detailed description of the task of automated musical accompaniment, and subsequently narrows the focus of the problem directly addressed in this work.

1.2 The Problem of Automated Musical Accompaniment

Performing as part of a musical ensemble requires more than just mastery of an instrument. Performers in an ensemble must listen to one another and respond to changes in the ensemble performance. These changes may include adjustments to tempo, dynamic, and articulation. Each performer in the group must be able to synchronize with the rest of the ensemble by playing at the appropriate time and with the appropriate tempo. The automated accompaniment problem is the task of enabling a computer to follow a performance by a solo musician or a group of musicians and to perform one or more voices of a composition in synchrony with the live performer or performers. It is assumed that all performers, including the computer, play from a score that indicates at least pitch and relative timing for every instrument over the entire piece.

The problem of automated accompaniment can be partitioned into four distinct subproblems:

1. Sensing and observing the performance ("Hearing").
2. Tracking the performer through the score ("Listening" or "Perceiving").
3. Adjusting parameters of the computer performance ("Control").
4. Generating the actual accompaniment performance ("Playing").

Any accompaniment system that attempts to be responsive to the live performer must address each of these tasks. In addition, the real-time requirements of musical performance demand efficient solutions to all of these problems. The system must execute these tasks concurrently and must react to the live performer in a manner that is aesthetically acceptable.

The first subproblem for an accompaniment system is detecting what the live musician has performed. The objective here is to extract from the performance the musical information that is important to tracking the performer and controlling the accompaniment. This information may include parameters such as fundamental pitch, note duration, attack (onset of a note), dynamic (relative loudness), and articulation. The exact parameters that a system extracts may vary according to the type of performance it tracks, how the performance is represented when it is input to the system, the expense of determining certain parameters, and the accuracy with which the parameters can be extracted. In a very straightforward case, digital messages sent from an electronic keyboard can provide highly reliable and precise information about pitch, duration, and dynamic. This information can be extracted with almost no processing because keyboards contain position sensors or switches for each key. In a more difficult case, extracting parameters by processing and measuring digitized sound signals requires more computation time and is likely to be less accurate.

The second task of an accompaniment system is tracking the performer's movement through the score. The system must generate estimates of at least the performer's current location in the score and rate of motion (tempo). The system must have this information in order to synchronize with the performer. Estimates of the performer's position must have high accuracy, precision finer than a note (*i.e.*, it matters whether the singer is near the beginning, middle or end of the note), and low latency (*i.e.*, must estimate the position as of a small fraction of a second ago, not the position several seconds ago). The method of generating position and tempo estimates must utilize both the parameters extracted from the actual performance and a score of the piece. In order to provide accurate estimates of position and tempo, the system's performance tracking method must be robust relative to both the inaccuracies of the parameter extraction process and the discrepancies between the literal score and the performer's rendition. Generally, the less accurate the position and tempo estimates are, the worse the system will be at synchronizing. Thus, lack of accuracy can drastically reduce the aesthetic value of the performance.

The third task of automated accompaniment is adjusting control settings for the performance in an appropriate manner. This task may include synchronizing with the performer's current position in the score, altering the tempo of the accompaniment, and adjusting volume. The particular control actions taken will be based on the estimates generated by the performance tracking system. As previously mentioned, these estimates include at least the score position and tempo of the performer. To select an aesthetically acceptable action, the accompaniment system must have some method for evaluating the appropriateness of possible actions within the context of the current state of the performance. As a simple example, suppose the accompaniment system is able to increase or decrease its tempo. If the tracking system's position estimates indicate that the performer is currently ahead of the accompaniment, then the system would ideally chose to increase its tempo in order to synchronize with the performer. In addition, the system must determine the magnitude of the tempo increase. This might require the system to simultaneously satisfy potentially conflicting requirements, such as minimizing the time to synchronization while maximizing the overall aesthetic of the performance (*e.g.*, not skipping notes in the melody).

The final task for accompaniment systems is generating the accompaniment. This task involves converting a symbolic representation of the accompaniment score into actual sound, according to the current settings for the performance control parameters (such as tempo). In the simplest case, this playback may constitute simply issuing commands to sound synthesizers. A more complex approach might actually generate digital sound samples for real-time output to a digital-to-analog converter in a sound card. Regardless of the exact sound production process, however, every accompaniment system must be able to satisfy the timing requirements indicated by the score. In addition, the specifics of the timing are likely to change during the performance as control parameters are adjusted. The dynamic nature of musical timing demands that the system apply some form of dynamic scheduling to sound production.

While there are many interesting component problems comprising the task of automated accompaniment, this work focuses on the problem of tracking a performer. Specifically, it addresses the task of accurately identifying score position and tempo of a vocal performer singing a composed piece of music. In addition, consideration is given to using the estimates of the performer's score position and tempo to alter the system's performance, hopefully enabling it to synchronize with the performer in an aesthetic manner. The tracking system is evaluated by incorporating it as part of a previously constructed accompaniment system (Grubb and Dannenberg 1994a; Grubb and Dannenberg 1994b).

The remainder of this chapter provides background on automated accompaniment. The next section provides a summary of previously constructed automated accompaniment systems, including those intended for acoustic instruments as opposed to vocalists. This historical overview is followed by a summary of the limitations and possible sources of error that affect the accuracy of the tracking methods used by these systems. Next, the difficulty of tracking vocal performers is discussed within the context of these possible sources of error. Specific examples are included in order to give the reader some intuition about why tracking a vocalist is problematic. This chapter concludes with a statement of a new approach to solving the vocal performance tracking problem. This approach is explored throughout subsequent chapters. The final section includes discussion of why this approach offers improved tracking of singers and in turn can lead to enhanced automated accompaniment for vocal performances.

1.3 Summary of Previous Work

Work on automated musical accompaniment began in the mid-1980's. Two seminal papers on the subject, one by Dannenberg (1984) and one by Vercoe (1984), appeared at the 1984 International Computer Music Conference. Vercoe's system was designed to track flute performances. Parameters extracted from the performance appear to have included pitch and duration information. In addition, optical sensors installed on the flute keys provided fingering information that helped to disambiguate pitch. The tracking technique appears to have applied a tiered approach to matching parameters with the score. The first order match simply compared performance parameters to the next expected note in the score. It indicated a match in sequence as long as the parameters and the score did not differ by "too much". If two or more notes in succession appeared not to match according to the heuristics applied, then some form of more extensive pattern matching was applied. Matches were used to estimate tempo at fairly frequent intervals. A form of smoothing was applied over longer periods of time. This "averaged" tempo estimate was used to synchronize the accompaniment to the soloist.

Dannenberg's system was initially applied to accompany performances on electronic keyboards, and subsequently performances on brass instruments. The only parameter extracted from the performance was pitch. In the case of acoustic instruments, pitch detection was accomplished exclusively through signal processing techniques. These techniques also attempted to segment the signal (*i.e.*, distinguish note transitions), so that most often pitch was reported only once for each performed note. In contrast to Vercoe's "in-sequence" approach, the tracking technique was based upon a dynamic programming algorithm for comparing two character strings. Ratings were accumulated for all possible series of matches between the sequence of extracted parameters and the sequence of notes in the score. A match between an extracted parameter and a note in the score added 1 to the rating, while skipping a note in the score between matches incurred a penalty of 1 against the rating. The last score position in the highest rated series of matches became the current estimated position of the performer. It was assumed that extraction of pitch occurred within a consistent period of time after the performer began the note. Tempo could thus be estimated by first associating each score position with the time at which a highest rated series of matches ended at that position, and subsequently calculating amount of score traversed per unit time over several successive matches.

During the decade following the publication of these papers, several extensions and embellishments to the basic techniques for score following were described. Vercoe and Puckette (1985) applied statistical methods for determining tempo changes that a specific performer could be expected to make at specific locations in the score. This information was acquired gradually through rehearsal with the accompaniment system. Meanwhile, Dannenberg and Bloch (1985) described an extension to the dynamic programming technique to enable tracking of polyphonic performances—performances where multiple notes are played simultaneously (such as in pieces for piano). Dannenberg and Mukaino (1988) presented heuristics for improved score following. These techniques are fairly independent of the underlying process that matches the performance to the score. They provide ways of dealing with ornaments and embellishments (such as grace notes and glissandi), which can be difficult to compare explicitly against a score. Dannenberg and Bookstein (1991) describe mechanisms for improving control and response of an accompaniment system. Although the described techniques are applicable to musical accompaniment, they detail experiences with a system intended for use by a conductor rather than a performer.

Another system for instrumental accompaniment was presented by Baird and her colleagues (Baird *et al.* 1989; Baird *et al.* 1993). Their method of score following rates a limited sequence of performed notes against short sequences in the score. This pattern matching approach considers duration of both notes and rests in addition to pitch, and it applies a rating system that allows for limited deviations between the performance and the score. The allowable deviations are based upon errors commonly made by either performers or parameter extraction methods. The length of the matched parameter sequences falls between the short sequences first considered by Vercoe's method and the longer sequences considered by Dannenberg's dynamic programming approach. This system heuristically incorporates more parameters and information per match than either of the initial systems, but the pattern matching approach also requires more computation.

Finally, Grubb and Dannenberg (1994a; 1994b) describe a method for tracking and accompanying a group of live musicians. This technique first applies the dynamic programming method to track each individual player in the ensemble. In order to estimate the position and tempo of the group as a whole, the individual score position and tempo estimates for all performers and the accompaniment are weighted and combined. The weights are based upon how recently new input has been received from a performer and how closely each individual position estimate corresponds to the other estimates.

The first system used to accompany vocalists applied a blackboard architecture with numerous production rules (Lifton 1985). Blackboard systems consist of a global repository (the blackboard), production rules that specify actions or calculations performed when certain preconditions are satisfied by information in the repository, and a scheduling component for controlling activation of the production rules. In Lifton's system, the repository contained a representation of the score and two parameters measured from the live performance—pitch and timing. Although extensive details of the implementation are not published, the system appears to have relied on significant hand-tailoring of production rules and customized expectations of duration and timing. These expectations were based on averages calculated during rehearsal of a specific piece with specific performers.

Several systems for tracking vocal performances have been more recently described. Katayose and colleagues (Katayose *et al.* 1993) and Inoue and colleagues (Inoue *et al.* 1993) describe two different systems for following vocalists performing contemporary pop music. These systems apply different methods and heuristics for determining pitch of the vocal performance, and appear to use "in-sequence" tracking techniques similar to Vercoe's. Both papers note difficulties in reliably extracting pitch from vocal performances, particularly when the continuous pitch information must be segmented and compared against a discrete score.

Subsequently, Inoue and colleagues (Inoue *et al.* 1994) presented a modified version of the system that relies on speech processing techniques to identify the current vowel in the performance. The tracking system compares the results of the vowel recognition against the score. This information appears to be more reliable than pitch in the case of amateur performances of pop music. Their system requires the singer to provide examples of sung vowels prior to performing with the accompaniment. In order to further reduce spurious recognition, the results of speech processing are weighted according to recency and are smoothed prior to comparing them against the score. This system also applies an interesting model of man-machine interaction when it adjusts the tempo. The model attempts to account for tempo expectations on the part of both the computer and the live performer when calculating the tempo changes needed to maintain synchronization.

Puckette (1995) also presents a system for vocal accompaniment. In contrast to the previous two systems, this software was used to accompany a contemporary art piece for computer and soprano. In addition, the piece did not apply such a strict sense of tempo as does the music targeted by the other systems mentioned. As a result, Puckette's system focuses on minimizing the delay in recognizing important transitions between notes (*i.e.*, those where the computer must react) rather than on estimating tempo. For purposes of score following, pitch is extracted from the performance. Puckette provides some heuristics for enhancing the in-sequence type of score following presented by Vercoe. These techniques include heuristics to deal with common problems such as vibrato, where the sung pitch is intentionally made to oscillate around the pitch written in the score. The position estimation also uses annotations that are placed in the score to indicate how reliably or unreliably certain notes can be detected during the performance. These markings provide specific "hints" about the reliability of the parameter extraction process relative to the piece and the performer.

Finally, the commercially available system from Coda Music Technology was extended in the mid-1990's for use by vocalists. Initially the system worked only with band instruments. Due to the real-world limitations on commercial products, including the allowable development effort and the need to run on very inexpensive personal computers, the tracking technology used for the vocal system does not seem significantly modified from the technology used to track instrumental performances. Tracking appears to be based primarily on fundamental pitch. The vocal product is also targeted for students, particularly younger and less experienced students who often work with easier repertoire and apply less elaborate expressive variation from the score. Coda's application also permits the user to annotate the score directly in a variety of ways.

1.4 Limitations of Previous Score Following Approaches

The challenge for any score-following system is to adequately deal with the variability of the musical information extracted from a performance relative to what appears in the score. Furthermore, tracking must be accomplished within the real-time constraints of the accompaniment task and must produce high accuracy, high precision position estimates. All the systems described in the previous section apply some form of rating scheme that is intended to encode and to implement heuristics for managing parameter variability. Several heuristics appear repeatedly across systems, even though the details of their implementations vary.

For example, all the systems have some method for preferring score locations within close proximity to previous estimates of score location, all other considerations being equal. This heuristic indicates that the system designers expect performers to render the score sequentially, and that more often than not the parameter extraction and comparison process is reliable enough to identify successive note transitions. Many systems incorporate some degree of leniency when rating matches between extracted parameters and the score. Few systems rate matches using an all or none approach, taking a strictly literal interpretation of the score. Some attempt to incorporate multiple parameters when rating matches with score locations; some use score annotations or expectations as part of the extraction process for a single parameter, in order to disambiguate the current value of that parameter. These heuristics indicate a general belief that using different kinds of information improves the accuracy of position estimates. A number of additional heuristics can also be identified.

The problem with these systems, however, is that the methods for rating matches and possible score locations vary significantly—even to the point where it is not possible to directly compare tracking systems. The way that various parameters and heuristics influence the position predictions can be quite different, depending upon both the rating system and the experiences of the designer. Tracking algorithms and ratings may be based on implicit assumptions regarding the extracted parameters, the style of music, or the instruments to be tracked.

The variation and subjectivity exhibited by these systems has several drawbacks. First, any detailed comparison of these tracking systems requires an understanding of multiple rating systems. It is difficult to apply to one rating method intuition about how another rating method deals with specific aspects of score following. Similarly, it is difficult to map heuristics that appear to work well for one approach onto another approach.

Second, the systems are not easily extended to incorporate new parameters or heuristics, or to accommodate changes in the methods used to extract parameters from a performance. The tracking systems are often targeted for specific parameters and their construction is often based largely on the designer's intuition. It is not easy to predict how a particular modification will affect accuracy of the tracking system without actually using the system to accompany live performances.

Third, it is challenging if not impossible to evaluate the systems in any analytically valuable way. For example, it is very hard to determine whether the failures of one system relative to another are due to either a truly inferior parameter extraction process or simply a poor rating scheme for comparing extracted parameters to the score. Because the rating systems and tracking algorithms are based upon heuristics, there is not an obvious way to evaluate whether or not a better rating scheme exists or what it would be.

Finally, to the extent that specific heuristics are applied within a similar context (*e.g.*, a strong preference for sequential score position matches when using pitch extracted from woodwind performances of Western classical music), a subjectivist representation and assessment of such heuristics seems questionable. One would expect that there is a certain quantifiable value to each piece of information relevant to score following. When estimating score position and tempo, all tracking systems should probably incorporate that information in a similar manner.

Expanding upon this latter point, one might expect that there exists a certain set of parameters or information that is useful for estimating score position of a performer and can be processed in real time by available computers. In the case of vocal performances, this information might include pitch, tempo, phonetics, breaths, rests, oral aperture, etc. In the best of all possible worlds, at most a couple or even just one of these parameters would be sufficient to disambiguate a performer's score position. In reality, however, the variation of the parameters relative to the score is significant, for a number of reasons. The following section presents some of the specific problems encountered when extracting parameters from vocal performances for use in score following. These problems, as well as the limitations of previous score-following systems, are used as motivation for developing a stochastic method of score following.

1.5 Problems with Tracking Vocal Performances

As a first attempt to construct a system for tracking vocalists, one might be tempted to match exact pitch information from the performances to the score. This approach to tracking, unfortunately, turns out to be less than adequate. Figure 1-3 contains a graph of the pitch extracted from the first phrase of a recording of "Happy Birthday" as performed by a university student majoring in voice performance. The pitch value was obtained by averaging the pitch period over frames containing 33 ms of sound signal, quantizing to the nearest semitone, and selecting the median pitch for every 3 successive frames. The variability of the pitch relative to the corresponding note in the score is quite apparent. Many of the notes were performed with *vibrato*, a trained vocal technique of varying the pitch in a periodic fashion. This variation would appear much more continuous if the pitch were not smoothed and quantized, but instead was plotted as frequency. Note also the initial upward slide of the pitch on the syllable "Hap-" at the beginning of the phrase. This slide results from an intentional stylization of the initial syllable by the performer.

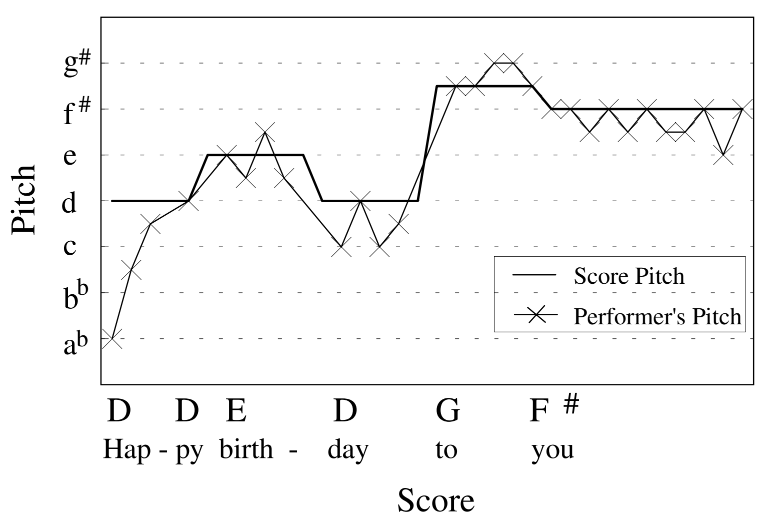


Figure 1-3. Performed pitch versus score pitch for a vocal performance.

In addition to this variability, the use of pitch information is limited by another property of the vocal scores. Table 1-1 contains statistics summarizing the syllable, word, and melodic interval content of the singer's part in each of twenty vocal scores randomly selected from a library collection. This collection consisted primarily of Western classical music, but contained a small amount of pop music as well. The table includes statistics for the percent of repeated pitches in the vocal melody, the percent of semitone steps, and the percent of whole tone steps. In the average melody, roughly 15% of the time a pitch is repeated from one note to the next. This fact implies that on average, a tracking system that uses only pitch information will be unable to distinguish 15% of the note transitions even assuming perfect extraction and matching. For the sample melody containing the most repeated pitches, a repeat occurred more than one-third of the time. Successive notes spanning a small melodic interval can also be difficult to distinguish since expressive pitch variations such as vibrato can be mistaken for a change of note. For the twenty examined melodies, on average over 30% of the intervals were either a semitone or a repeat, and more than 60% were a whole tone or less—within potential vibrato range. In general, and as might be expected, stepwise melodic motion appears to dominate vocal music.

As indicated in the first two columns of the table, using phonemes, syllables, or words alone is not likely to provide any better tracking than using pitches, even assuming improved extraction and matching. On average, only about 85% of the notes in a score are associated with a distinct syllable. The remaining 15% are simply a change of pitch sung on the same vowel. In the worst case found in the sample set of scores, less than 60% of the notes are associated with a change of syllable. The situation of course degrades when words instead of syllables are considered. In one of the examined scores, less than 40% of the notes are associated with a new word.

One might consider trying some simple segmentation technique to identify note transitions. This segmentation could be used in combination with either pitch or phoneme recognition aligned to the

Table 1-1. Summary statistics for syllable, word, and melodic interval content of the vocalist's part in each of twenty randomly selected vocal scores.

	# Syllables / # Notes	# Words / # Notes	Percent Repeats	Percent Semitones	Percent Whole Tones
Mean	0.846	0.578	15.4	19.2	30.9
Median	0.882	0.553	15.4	17.0	30.3
Minimum	0.588	0.366	3.8	4.0	7.0
Maximum	1.000	0.796	34.4	50.0	54.5

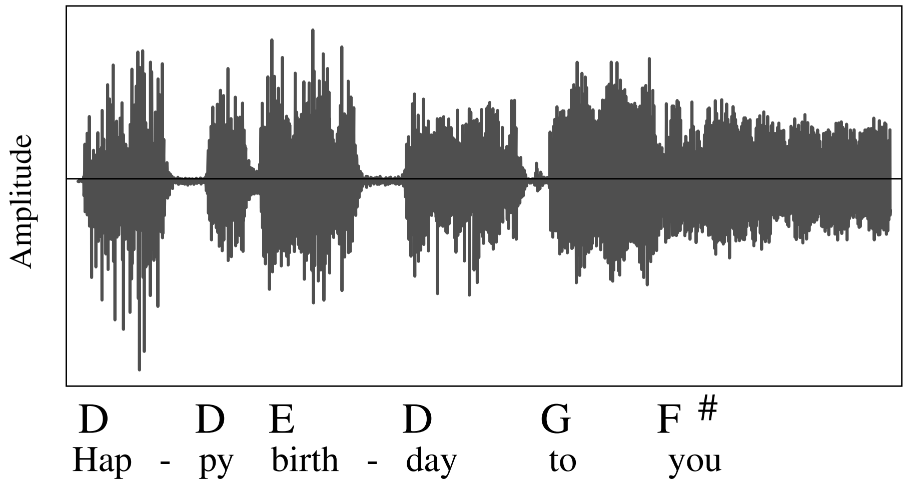


Figure 1-4. Sound waveform during a single phrase from a vocal performance.

transitions. Changes in amplitude, power, or noise content of a signal can be indicative of a change of note for vocal as well as instrumental performances. Unfortunately, as is shown in Figure 1-4, vocal performances can be extremely fluid and in many cases will not provide sharp and consistent parameter changes at note transitions. Consider the final note transition in the figure, between the words "to" and "you". Pitched sound of significant amplitude persists through the transition. The pitch itself is also likely to change in a smooth and continuous fashion. As indicated by various portions of the graph, changes in power and amplitude may occur within a note as well. This variation may result from a change of phoneme as well as an intentional change of volume for expressive purposes.

In summary, there exist several difficulties in trying to follow vocal performers by using information extracted from their performances. Parameters obtained from vocal performances exhibit high variability relative to what is written in the score. In addition, a single parameter (such as pitch or phoneme) often remains constant in the score through several successive notes, thereby limiting the utility of each individual parameter. Vocal performances are also very fluid and continuous, and therefore often fail to exhibit sharp and consistent parameter changes at note transitions. Consequently, it is difficult to reliably segment a vocal performance into notes. Finally, it is not uncommon to encounter sections of a vocal score where either the performer is expected to freely and sometimes suddenly adjust tempo, or where expressive timing may be applied within a consistent tempo. Unless the current score position of the performer can be accurately and precisely identified, such adjustments to the performance may go undetected by the accompaniment system. All of these problems conspire to make accurate identification of a vocal performer's score position an extremely difficult task. Robust score following for vocal performances will require integration of at least several different information sources.

1.6 Stochastic Score Following for Automated Accompaniment

Estimating the score position of a singer with high accuracy, high precision, and low latency is a difficult problem. The previous sections have described several specific reasons why vocal performance tracking is hard. Not the least of these reasons is that live musicians appear to consider multiple parameters or measurements of the sound signal, such as pitch, diction or phonetic content, silences and note onsets, dynamic level (amplitude), and tempo (rate of change). However, when considering how to construct software to track a vocalist, all of the mentioned problems can be summarized by noting that they all produce a single important effect—they hinder specification of a deterministic function that maps the available information (including the time at which measurements are taken) into the score position of a singer. In other words, they prevent system designers from specifying a complete one-to-one mapping from the set of available information (including a given score plus any sequence of extracted performance parameters) into the score position of the singer.

In general, inability to apply a complete, deterministic function may occur for a variety of reasons. Even when the mapping from the actual, real-world input values to the estimated values is a true function, use of this function may not be possible. Unreliable detection or measurement of the input values can hinder direct application of the function. For instance, reliance on analog sensors or measurements from analog signals may yield inaccuracies. Limits of available computation also may prevent use of a functional specification. If only limited signal processing and computation over a few inputs can be accommodated, the relation mapping those inputs to the target values may not be a true function. Even assuming the available inputs are accurate and limited compute power is not a problem, lack of knowledge about the function can make complete specification impossible. Such ignorance can result from not knowing all the inputs necessary to define a true function, inability to measure or obtain the values of some necessary inputs, and inability to fully specify the function because it is either very complex or highly arbitrary. For instance, vocal performers do not follow a simple set of rules when rendering a musical score, and accompanists apply significant musical knowledge and experience to follow these performances. Finally, it is always possible that no complete functional specification can be given, regardless of how many inputs are available, how accurate they are, and how much knowledge is obtained about the problem. This situation certainly may arise when attempting tasks that involve human perception and response. Since human accompanists are not always completely synchronized with the singer, it is possible that they sometimes do not have sufficient information to unambiguously determine the singer's score position. The accompanist cannot always infer what the singer is thinking.

As a way of dealing with the difficulties of score following for vocal performances—rating possible score positions of the performer, dealing with variability of extracted parameters, integrating multiple sources of information, analyzing performance of a tracking system, etc.—the development of a stochastic technique for score following is proposed. Rather than insisting on a deterministic function specifying position of the singer for all possible value assignments of the inputs, this approach will develop a probabilistic description of the singer's position given the available information. Instead of a subjective rating system for assessing possible performer score positions, this method will use probabilities indicating the likelihood that the performer is near a particular score position. The needed statistics may be estimated by experiments and analysis of data, allowing a tracking system to base its position estimates upon actual, observed behavior. Properly defined probabilities inherently may account for variability of the parameters extracted from a performance, as well as for uncertainty in the other information relevant to determining location. Combining multiple sources of information can likewise occur in a sound fashion, based upon the proper methods for combining probabilities. Finally, some analytic techniques may be applied to estimate statistics when comprehensive data is not available or is extremely difficult to obtain. Analytic techniques also could be applied to evaluate or predict performance of a tracking system, possibly to indicate how modifications to the system are likely to influence its ability to track a singer.

A stochastic model for score following is presented in the next chapter. Subsequent chapters consider the specific requirements for applying this general model to track vocal performances of Western classical music sung by trained singers. Finally, several completed tracking systems are described, each system applying a different set of information. Details are given as to how the tracking systems are incorporated within an automated accompaniment system. These different tracking systems are evaluated by using them to accompany both recorded and live vocal performances. This work demonstrates the viability of using a statistical model for vocal performance tracking and examines how combining multiple sources of information, including several components of the score and several measurements of the performance sound signal, improves tracking accuracy. In addition, it serves as a case study in applying statistical modeling in an incremental fashion to a problem for which a complete and adequate functional specification in any form is either unknown or nonexistent. Emphasis is placed both on the appropriate methodology for such incremental modeling and on assessing the accuracy and validity of the developed model.

Chapter 2

A Model for Stochastic Score Following

2.1 Motivation

Any system that attempts to track a vocal performer must deal with the uncertainty inherent in the estimates of the performer's score position. The real-time data extracted from vocal performances is generally noisy and inconsistent. This variation occurs for numerous reasons. The software analysis of the digitized sensor output can be unreliable, or the analog sensors themselves may be noisy and error-prone. Furthermore, the generative source of the signal is neither reliable nor consistent. Performers frequently deviate from a literal interpretation of what is written in the score, sometimes for extended periods of time. Such deviations may be intentional and even rehearsed, but sometimes they are spontaneous or accidental. However, in all cases these alterations are not explicitly notated in the score, and frequently they are difficult if not impossible to precisely characterize and reliably anticipate.

Several systems for vocal performance tracking and accompaniment have been previously described. While some of these systems have been "customized" for a particular performance or performer, others have been proposed for more general-purpose application and are intended for use by many singers performing numerous pieces. These latter systems, however, have met with limited success. Poor tracking can result from their inability to account for the extreme unreliability and variability of the information they use. As a result, they may not adequately manage the inherent uncertainty in their estimates of the performer's score position. Addressing this deficiency would likely enhance the musical performance of these systems, or at least provide a more precise characterization of the uncertainty in the score position estimates they generate.

In response to this problem with score-following systems, a model for stochastic score following is presented. Starting from a fairly loose definition of a score, a probabilistic description of a performer's score position is developed. This idea is extended by presenting a general model of how the stochastic description changes over time, based upon relevant information that can be extracted from the vocal performance in real time. The general model is subsequently modified according to a limited number of fairly mild assumptions. These assumptions permit development of a final model that is tractable from the standpoint of both computation and statistical estimation. Finally, a description of a software implementation of the model is provided, along with an analysis of the error introduced by a discrete approximation to the continuous model.

2.2 Definition of a Score

The concept of a *score* (and specifically a musical score) will be used as a foundation for all subsequent stochastic modeling and analysis. For our purposes, a *score* will be loosely defined as a sequence of *events* that is based upon a "strong" expectation of a desired or fixed ordering. Each *event* is specified by:

1. A *relative length*, expressed in some unit of distance or time, that defines the size or duration of the event relative to other events in the score.

2. A *distribution*, expressed as either a probability function or a density function, that completely specifies the observation probabilities (the likelihood of every possible sensor output) at any time during the event, and is conditioned on at least a simple classification of that event.

Several aspects of this definition are worthy of further discussion and clarification.

The term *event* is used throughout this discussion. It will always refer to a single probabilistic characterization of the possible sensor output that is observed whenever a performer is within the corresponding region of the score. It will always refer to a probabilistic characterization based upon the symbolic representation contained in the score, as opposed to specific real-world performances or renderings of that portion of the score. The term *performed event* will be used when referring to the latter.

It may be momentarily helpful for the reader to associate an event with a single note in a musical score. The type of note (*i.e.*, quarter, eighth, sixteenth) would indicate a relative length for the event. A probabilistic characterization of this note might consist simply of a set of probabilities indicating the likelihood of observing a particular musical pitch as output from a pitch detection device. If the pitch of the note in the score were an A-440, then A-440 and notes close to it would likely have higher probability of appearing as output from the pitch detector than would notes more distant in pitch. It is important to point out that in this example, the distribution associated with each event in the score is conditioned on pitch. The events found in the score therefore are classified according to the pitch indicated in the score.

While this example of an event may help to ground the more abstract definition, the reader should not assume that the term *event* is synonymous with a musical note and its associated pitch probabilities. Several different examples of events will be examined, and the formal models for score tracking do not rely on specific assumptions regarding the types of sensor output or the symbolic representation used in a score. Several chapters are devoted both to detailing various types of events for characterizing musical scores and to analyzing how different definitions of events may affect the tracking performance of the stochastic model.

A score is specified as an ordered sequence of events, each event having a "fixed" position relative to the other events in the score. Any repeated sections in the score are expanded to yield a simple linear sequence of events. The position of each event is said to be "fixed" because it is believed with high confidence that, for the most part, the ordering of the events will be preserved during performance. In addition, it is believed with some degree of confidence that the length or duration of each event will be maintained relative to the lengths of other events within close proximity. For example, if the first of two consecutive notes in a score is a half note and the second is a quarter note, then it is expected that in a performance of this score, the duration of the first note is likely to be roughly twice that of the second note. This expectation preserves the two-to-one ratio indicated in the score.

The term *length* may seem awkward or inappropriate when discussing events contained in a musical score. Notes and larger sections of a score generally are thought of as having a relative, idealized duration in time, rather than spanning a length or distance. However, it will be necessary at various points in this document to discuss the *actual* duration in time of a note as it is performed by a soloist, rather than the *idealized* relative duration of that note as indicated in the score. In an attempt to minimize confusion, the terms *note duration* and *performance time* will be used when referring to the former type of duration, whereas the terms *note length* and *score distance* will be used when referring to the latter type of duration. It is appropriate to think of the terms *time* and *duration* as referring to the actual passage of time resulting from a rendering of the score, while the terms *length* and *distance* will refer to the idealized dimension characterizing the spacing of events in the score.

Many aspects of the given definition of a *score*, including length of an event, have been discussed as if the values actually associated with these parameters are likely estimates rather than fixed and invariant values. This approach will carry through to the formal model proposed for score tracking. In the latter case, however, the nature and extent of such uncertainty will be made more precise. The accurate characterization of this uncertainty and its propagation over time will be a primary concern when defining the general score-tracking model, as well as when subsequently applying that model to actually track musical performances.

It should be pointed out that, as presented here, the definition of score imposes a view in which both a score and the events in it are continuous. The events in a score are defined as having a real-valued length. Events are not viewed as discrete states that cannot be further divided. A score, in turn, is composed of a contiguous sequence of these events. It therefore makes sense to discuss position, both within an event and within a score, as being real-valued. The continuous nature of score position is important not only because of the implications it has for the subsequent theoretical model of score following, but also because of pragmatic requirements of musical score-tracking systems. The resolution of score position estimates made by a tracking system affects the performance of the accompaniment. Because a single note in a soloist's part may often extend over several notes in the accompaniment, a score-tracking system must either guarantee that all position estimates accurately refer to the same point within an event (*e.g.*, the singer has just started the note) or be able to estimate a soloist's score position at a resolution finer than a single event. The score-tracking model described here will attempt to do the latter by building upon the continuous nature of the definition of a score, as it has been presented in this section.

2.3 Definitions and Notation

Before pursuing a description of a performer's score position, and subsequently a model for defining and updating that description, an overview of some basic probabilistic notation and definitions will be given. This section is not intended to provide a comprehensive introduction to continuous probability, but simply to give a consistent, up front summary of the symbols and definitions that are used throughout this chapter and subsequent chapters.

Most of the theoretical material in this work deals with continuous probability, where the range of values for a random variable, X, is over a continuous region. Since these regions are continuous, it is not possible to discuss the probability of the variable assuming a specific value (as there are an infinite number of such values). Instead, we will consider the probability that the value assumed by a variable falls within a particular region of the range of all possible values. For example, the probability that a variable X assumes a value between a and b will be notated as $\mathrm{P}[a < X < b]$. In addition, each variable will have associated with it a *cumulative distribution function*, or *cdf*. This function defines the probability that the random variable assumes a value at or below a given point. For example, the cdf of X will be notated as $\mathrm{F}_X(x)$, and $\mathrm{F}_X(a) = \mathrm{P}[X \leq a]$. The cdf of a random variable is nondecreasing over the defined range of the random variable. In addition, the probability that a variable assumes a value within the region a through b can be defined in terms of the cdf as $\mathrm{P}[a \leq X \leq b] = \mathrm{F}_X(b) - \mathrm{F}_X(a)$.

The *density function* of a random variable is defined as the derivative of the variable's cdf, and will be notated as:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

The probability that a variable will assume a value in the region *a* through *b* can be defined in terms of the definite integral of the density function over the region *a* through *b* :

$$\mathrm{P}\,[\, a \le X \le b \,] = \int_a^b f_X(x)\, \partial x$$

Graphically, this probability is represented by the area under the density function between $x = a$ and $x = b$.

A *multivariate density function* is a function of several variables such that a definite integral of that function over all variables gives the probability that all variables will simultaneously assume a value within that multidimensional region. For example:

$$\mathrm{P}\,[\, a \le X \le b \,,\, c \le Y \le d \,] = \int_a^b \int_c^d f_{XY}(x, y)\, \partial y\, \partial x$$

where $f_{XY}(x,y)$ is a joint density function. A *marginal density function* is a density function over only a few variables from the joint density, and can be obtained by integrating the joint density over the entire range of all other variables:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)\, \partial y$$

Following the relationship between the density function and the cdf, the multivariate density can be integrated over all variables to yield the *multivariate cdf*. Likewise, the marginal density can be integrated to yield the *marginal cdf*.

A *conditional density function* is the density of a variable *X* given one or more variables, and is defined as the ratio of the joint density of the variables and the marginal density of *X*, wherever the latter density is positive:

$$f_{X|Y}(x|\, y) = \frac{f_{XY}(x, y)}{f_X(x)}$$

and zero otherwise. Note that this equation can be rewritten to express the joint density as a product of the conditional density and the marginal density. This product can be substituted in place of the joint density function in the previously described integrals for calculating probabilities, other marginal densities, and cdf's.

It is important to point out that for fixed values of the conditioning variables (the variable Y in the above example), a conditional density is a true density function over the conditioned variable (the variable X in the above example). In other words, substituting values for all conditioning variables and integrating the resulting function over the entire range of the conditioned variable will yield a value of 1:

$$\int_{-\infty}^{+\infty} f_{X|Y}\,(x \mid y = a)\,\partial x \;= 1$$

The equivalence notation used inside of the conditional density function in this example will appear from time to time throughout this work. It will always be used to indicate that a fixed, known value for a conditioning variable has been substituted into a conditional density function that actually is defined over the entire range of values possibly assumed by that variable.

2.4 A Probabilistic Description of Score Position

A score consists of an ordered sequence of events. Each of the events, as well as the overall score, has an associated real-valued length. One possible characterization of a performer's position within such a score would be a single, real-valued point within the length of the score. However, this simple characterization alone does not afford a probabilistic description that accounts for uncertainty in the method of estimation. This approach does not assist us in achieving a representation of score position that is any better than those used by previously constructed accompaniment systems. These systems generally used the onset time of the closest note in the score as determined by their "best guess".

Our characterization of a performer's score position will instead be a continuous density function over the length of the entire score. This function will be referred to as a *score position density function*. The area under this curve between two points in the score will indicate the likelihood that the performer is in that continuous region of the score. As an example, consider the score presented in Figure 2-1. A density function indicating the likelihood of the performer's score position is drawn above the score. The

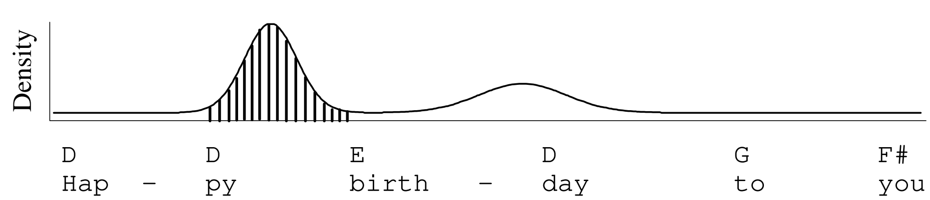


Figure 2-1. Example of a density function characterizing the score position of a performer. The area of the shaded region gives the probability that the performer is singing the second note.

shaded region of this function indicates the probability that the performer is between the second and third notes of the score (*i.e.*, is currently singing the second note). Since the function is a density function, the integral of the function over the entire score will always be 1. This property corresponds to assuming that the performer is always somewhere in the represented score.

This characterization of a performer's score position permits us to compare the probability that a performer is in one region of the score against the probability that the performer is in some other region of the score. Since the area under a region of the density function indicates the probability that the performer is in that region, the ratio of the area under the curve in one region to the area under the curve in a second region defines the *odds* that the performer is in the first region versus the second region. The region more likely to span the performer's location is the one having the larger area. Figure 2-2 provides a graphical example of this. The concept of odds will be helpful when defining the decision-making process that an accompaniment system can use to control its own musical performance.

While the use of a density function to characterize a performer's current score position is fairly straightforward, there are many questions regarding how this function is determined, how it is updated to reflect the changing position of the performer, and how it can be applied to make a decision about accompaniment control. The answers to these questions must be twofold, providing both the precision and consistency needed for a useful theoretical model while simultaneously permitting an efficient software implementation that satisfies real-time processing requirements without sacrificing accuracy. In addition, a score-following system must incorporate a variety of information relevant to accurate score position identification. This information includes the types of sensors and the specific observations they provide, the representation of the score and the events it contains, and the tempo at which a musician is performing. The statistical modeling and representation of this data, as well as any relevant errors or uncertainty, will have significant impact on the accuracy of the tracking system.

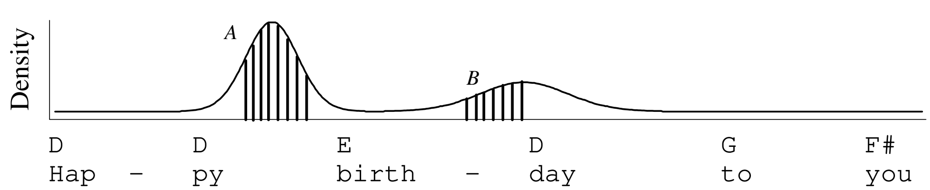


Figure 2-2. If A represents the area of the leftmost shaded region and B represents the area of the rightmost shaded region, then A / B gives the odds that the performer is in the first region versus the second region.

2.5 A General Model for Stochastic Score Following

The score position of a performer at any point in time is defined not as a single point, but instead is described stochastically by a density function over the length of the score. The nature of this function and its evolution during a performance will be influenced by many factors. These factors include the events that compose the score, their probabilistic description, the sensors used to observe the performance, and characteristics of the performance such as the performer's tempo and interpretation of the score. The purpose of this section is to define a very general model of the score position density function. This model will incorporate all available information relevant to estimating the performer's score position, and will provide a basis for updating the score position density over time.

While knowing the general shape and other characteristics of the score position density function may help us to model the performer's motion over time, no specifics about the function will be assumed in this chapter. For present purposes, it is only assumed that the function is a true density function over the score and therefore assumes a nonnegative value everywhere it is defined, is piecewise continuous, and integrates to 1 over the length of the score. The model of the score position density presented in this chapter consists of several component density functions that can be modeled directly. A more precise definition of these functions, for purposes of following a musician, is the topic of several subsequent chapters.

The model presented here is intended to support real-time score tracking for an automated accompaniment system. A high-level, operational overview of this process can be given as follows. There is some initial stochastic estimate of the performer's score position, characterized by a density function over the score. After a certain period of time, input is received from one or more sensors that extract parameters relevant to a musical performance. Inputs might include information such as fundamental pitch of a sound signal. At this point, the accompaniment system must update its stochastic estimate of the performer's score position, making use of any available, relevant information (such as the sensor output). It must then use this updated position estimate to determine what if any action should be taken to adjust performance of the accompaniment. This process will be repeated as new sensor output is generated, presumably at a fairly frequent interval, until the performance is completed.

The stochastic model for score tracking incorporates three kinds of information relevant to predicting the position of a performer. First, when generating a new score position density function, it incorporates the prior score position density function generated at the time the last sensor input was received. This function will be referred to as an estimate of the performer's *source position*. For distinction, the current position of the performer at the time the new sensor input is received will be referred to as the performer's *destination position*. The general model discussed here provides a method for generating an estimate of the performer's destination position.

Second, the stochastic model must consider any and all data just recently received from sensors that extract parameters from the performance. This data will be referred to as the *observations* made by the accompaniment system. Assuming that the sensors are fairly accurate and reliable, and focus on parameters relevant to identifying the events in the score, this information will be extremely important to score position identification.

Finally, the stochastic model must consider the motion of the performer through the score during the elapsed period of time. This information may include both direction and amount of score traversed. Since the system cannot know this information exactly, it will need to estimate this distance, possibly based in part on an estimate of the performer's tempo and the elapsed time. This characterization of the performer's motion through the score will be referred to as the *estimated traversed distance*, or simply the *estimated distance*. The exact manner of estimating this value for purposes of musical score following is not crucial to development of the general model, but will be discussed more thoroughly in a subsequent chapter. It is important to note, however, that the estimated distance is in fact an estimate, and the general model is not based on any assumptions regarding the accuracy or precision of this estimate. This uncertainty is represented explicitly by density functions appearing within the general model.

Now ideally, the model for score following should take all of this information—the performer's source position, the observations, and the estimated distance—and define a density function over the destination position. This function is simply a conditional density, and we will notate it as follows:

$$f_{I \mid D,V,J}(i \mid d, v, j)$$

i = the performer's destination position

d = the estimated distance

v = the observation(s)

j = the performer's source position

where each of the variables is defined as previously discussed. While the observation variable, v, is treated throughout this chapter as a scalar value, there is no reason why it cannot represent a vector of values, with appropriate modifications to the dimensionality of the density functions containing v. In this way, the presented model can be extended to deal with multiple observations (of possibly different parameters) simultaneously reported from the same sensor or different sensors.

This conditional density function defines the probability P[$I \leq i$] for any set of fixed values for the variables d, v, and j. Now in situations where these three variables assume fixed values, the conditional density (if we could define it) would provide all the information we needed. In reality, however, we know that at least the source position, j, is not known with certainty. The density function characterizing the source position is in fact the function that results from the previous iteration of our proposed model. What we would really like to have is a density function over the destination conditioned on only the estimated distance and the observation. By the definition of conditional density, such a function is obtained from our original conditional density as follows:

$$f_{I \mid D,V}(i|\ d,\ v) = \int_{j=0}^{\|Score\|} f_{I \mid D,V,J}(i|\ d,\ v,\ j) \quad f_{J \mid D,V}(j|\ d,\ v)\ \partial j \qquad \text{[2.1]}$$

where $\|Score\|$ represents the length of the score. This equivalence can easily be established by first applying the definition of conditional density to transform equation 2.1 into:

$$f_{I \mid D,V}(i|\ d,\ v) = \int_{j=0}^{\|Score\|} \frac{f_{I,D,V,J}(i,\ d,\ v,\ j)}{f_{D,V,J}(d,\ v,\ j)} \cdot \frac{f_{D,V,J}(d, v,\ j)}{f_{D,V}(d,\ v)}\ \partial j \qquad \text{[2.2]}$$

By canceling identical terms in the numerator and denominator, and moving the function $f_{D,V}$ outside of the integral, equation 2.2 becomes:

$$f_{I \mid D,V}(i|\ d,\ v) = \frac{1}{f_{D,V}(d,\ v)} \int_{j=0}^{\|Score\|} f_{I,D,V,J}(i,\ d,\ v,\ j)\ \partial j \qquad \text{[2.3]}$$

The integral in equation 2.3 is over a single variable from a multivariate joint density function. The integral evaluates to a multivariate marginal density, and thus equation 2.3 can be rewritten as:

$$f_{I \mid D,V}(i|\ d,\ v) = \frac{f_{I,D,V}(i,\ d,\ v)}{f_{D,V}(d,\ v)} \qquad \text{[2.4]}$$

which is the definition for the conditional density.

If we assume the density $f_{J|\ D,V}$ to be equivalent to the previous estimate of the performer's score position, then equation 2.1 can be rewritten as:

$$f_{I \mid D,V}(i|\ d_{n+1},\ v_{n+1})_{n+1} = \int_{j=0}^{\|Score\|} f_{I \mid D,V,J}(i|\ d_{n+1},\ v_{n+1},\ j) \quad f_{I \mid D,V}(j|\ d_n,\ v_n)_n\ \partial j \qquad \text{[2.5]}$$

To distinguish the conditional density that provides an estimate of the performer's previous score position from the one that provides an estimate of the performer's current score position, a subscript indicating sequence of evaluation is added. Similar subscripts are added to the variables representing the estimated distance and the observation.

It is important to note that this last substitution of functions requires that the stochastic estimate of the source position, j, either is independent of the current observation, v_{n+1}, and the current estimated distance, d_{n+1}, or already accounts for the current observation and the current estimated distance. Since neither the method by which observations are generated nor the method of estimating the distance traversed by the performer has been specified, it is at this point not invalid to make such an assumption. However, once these methods have been defined for purposes of musical score following, it will be necessary to consider whether or not this substitution remains valid. If the substitution is not entirely valid, it will be necessary to examine how this substitution will affect the density function generated by the model relative to the density that would be generated by a more accurate model.

Now, equation 2.5 presents us with a general model for stochastic score following based upon a source location of the performer, an estimate of the score distance traversed by the performer since the last observation was reported, and a new observation reported by a sensor (or sensors). In order to simplify the notation of this expression for future reference, the function notated as $f_{I\,|\,D,V}()_n$ will be notated as f_{Prior} and all evaluation sequence subscripts will be dropped, giving:

$$f_{I\,|\,D,V}\,(i|\;d,\;v) = \int_{j=0}^{\|Score\|} f_{I\,|\,D,V,J}\,(i|\;d,\;v,\;j)\, f_{Prior}\,(j)\;\partial j \qquad [2.6]$$

While this model is fine from a high-level point of view, it does not submit readily to a practical implementation. The density $f_{I\,|\,D,V,J}$ is a complex multivariate function that would be extremely difficult to explicitly define over any reasonable ranges for the variables. The difficulty would be compounded if any interesting dependencies existed amongst the variables. In addition, the evaluation of the definite integral could be fairly time-consuming depending upon how this function is defined. To address these problems, the next several sections apply some fairly general but simplifying assumptions in order to transform equation 2.6 into a more pragmatic model. This simplification is accomplished so as not to preclude a reasonable characterization of important interdependencies amongst the variables in the density functions.

2.6 An Independence Assumption

The conditional density inside of the integral in equation 2.6 would be difficult to explicitly define. It is helpful to rewrite this density function as a combination of density functions that could be defined more easily on an individual basis. As a first step in this direction, the definition of conditional density allows us to rewrite the conditional density from equation 2.6 as a ratio of the joint density and a multivariate marginal:

$$f_{I\,|\,D,V,J}\,(i|\;d,v,j) = \frac{f_{I,D,V,J}\,(i,\;d,\;v,\;j)}{f_{D,V,J}\,(d,\;v,\;j)} \qquad [2.7]$$

In addition, the definition of conditional density further allows replacement of the joint density by a product of a marginal and a conditional:

$$f_{I\mid D,V,J}\,(i\mid d,v,j) = \frac{f_{I,D,J}\,(i,\ d,\ j)\ \cdot\ f_{V\mid I,D,J}\,(v\mid i,\ d,\ j)}{f_{D,V,J}\,(d,\ v,\ j)} \quad \text{[2.8]}$$

Now in order to simplify the right-hand side of equation 2.8, we will make an assumption that in reality is probably not entirely justifiable. Namely, we will assume that the current observation, *v*, is completely independent of the estimated distance, *d*, and the source position, *j*. This assumption is made purely for the purposes of defining a tractable model and allows equation 2.8 to be transformed into:

$$f_{I\mid D,V,J}\,(i\mid d,v,j) = \frac{f_{I,D,J}\,(i,\ d,\ j)\ \cdot\ f_{V\mid I}\,(v\mid i)}{f_{D,J}\,(d,\ j)\ \cdot\ f_V\,(v)} \quad \text{[2.9]}$$

The definition of conditional density allows a further simplification of equation 2.9 to give:

$$f_{I\mid D,V,J}\,(i\mid d,v,j) = f_{I\mid D,J}\,(i\mid d,\ j)\cdot\frac{f_{V\mid I}\,(v\mid i)}{f_V\,(v)} \quad \text{[2.10]}$$

Substituting the right-hand side of equation 2.10 for the conditional density inside the integral in equation 2.6 permits:

$$f_{I\mid D,V}\,(i\mid d,\ v) = \int_{j=0}^{\|Score\|} f_{I\mid D,J}\,(i\mid d,\ j)\cdot\frac{f_{V\mid I}\,(v\mid i)}{f_V\,(v)}\cdot f_{Prior}\,(j)\,\partial j \quad \text{[2.11]}$$

In addition, moving the ratio outside of the integral gives:

$$f_{I\mid D,V}\,(i\mid d,\ v) = \frac{f_{V\mid I}\,(v\mid i)}{f_V\,(v)}\cdot\int_{j=0}^{\|Score\|} f_{I\mid D,J}\,(i\mid d,\ j)\cdot f_{Prior}\,(j)\,\partial j \quad \text{[2.12]}$$

It should be noted that if f_{Prior} is assumed to be equal to $f_{J\mid D}$, then the integral really just computes the density function $f_{I\mid D}$.

A few comments regarding the validity of the independence assumption are appropriate at this point. First, it is probably not true in general that the observation reported by a sensor is independent of the source position of the performer. If a sensor such as a pitch detector is processing a continuous signal, then it is likely that the signal will vary according to the source position of the performer. However, if the most likely source positions for a given destination position are associated with very similar signals (*i.e.*, the distributions of the observation are very similar), then this independence assumption may still permit a reasonably accurate model. This situation actually may be the case for score following where there is a high expectation that the score will be traversed in sequence.

Second, it is also likely that the observation reported by a sensor is not independent of the estimated distance, particularly if it is believed that the estimate is often a close approximation of the actual score distance covered by the performer. For example, if a sensor is effectively applying some sort of "centering" to a signal parameter (such as effectively calculating some form of average value or median value, perhaps average or median pitch) over a portion of a signal, then the amount of score traversed by a performer may affect the value reported by the sensor. However, such influences can be minimized by guaranteeing that the amount of signal analyzed per observation corresponds to a fairly small and consistent portion of score. This situation will often be the case if observations are reported so frequently that only a small portion of signal (relative to the signal's rate of "variation") contributes to each observation. Thus from a pragmatic viewpoint, the errors introduced by the independence assumption may be fairly inconsequential and well worth the increased potential to simplify both the implementation of the model and estimation of the component density functions.

A consideration of the density functions in equation 2.12 reveals that they in fact are likely to be more easily specified than the conditional density in equation 2.6. $f_{I\,|\,J,D}$ is a three-dimensional function based exclusively on score positions and estimated distance. Since it only considers position and distance, one can be hopeful that such a density could be empirically measured if not formally derived. $f_{V\,|\,I}$ is a function of sensor output and a single score position, so investigating the output of a chosen sensor at various points in the score (which hopefully correlates with what is *written* in the score) should provide a reasonable estimate of this function. Finally, if we allow the density f_V to be rewritten as the integral over i of the product of $f_{V\,|\,I}$ and $f_{I\,|\,D}$ (since $f_V = f_{V\,|\,D}$ and $f_{V\,|\,I} = f_{V\,|\,I,\,D}$ under the independence assumption), equation 2.12 becomes:

$$f_{I\,|\,D,V}\,(i|\ d,\ v) = \frac{f_{V\,|\,I}\,(v|\ i) \cdot \int_{j=0}^{\|Score\|} f_{I\,|\,D,J}\,(i|\ d,\ j) \cdot f_{Prior}(j)\ \partial j}{\int_{i=0}^{\|Score\|} f_{V\,|\,I}\,(v|\ i)\ \cdot\ \left[\int_{j=0}^{\|Score\|} f_{I\,|\,D,J}\,(i|\ d,\ j) \cdot f_{Prior}(j)\ \partial j\right] \partial i} \qquad \text{[2.13]}$$

Equation 2.13 contains only the three density functions, f_{Prior}, $f_{I\,|\,D,J}$, and $f_{V\,|\,I}$, that can feasibly be estimated in an explicit form.

It should also be noted that this independence assumption usurps some of the burden for the substitution of f_{Prior} for $f_{J\,|\,D,V}$ that occurred during development of the general model. If the current observation is stochastically independent of both the prior score position and the estimated distance, then $f_{J\,|\,D,V} = f_{J\,|\,D}$. Thus f_{Prior} in equation 2.13 really substitutes for $f_{J\,|\,D}$. To the extent that one accepts the arguments provided in this section as to why such an independence assumption is valid, one will be less dissatisfied by the general model for stochastic score following. Any subsequent discussion of problems caused by using f_{Prior} in the model will focus on consideration of how well this function approximates $f_{J\,|\,D}$.

As previously pointed out, this will depend in part upon the method for generating an estimated distance. Details of this will be presented in Chapter 3.

2.7 An Assumption of Fixed Conditioning Variables

While equation 2.13 provides a model for which it is tractable to define the component density functions, there are some potential problems regarding real-time computation of the density $f_{I\,|\,D,V}$. There remain a total of four conditioning variables throughout the density functions in the model. While one variable has been removed from the final density through integration explicitly notated in the model, two of the remaining conditioning variables carry through. If the estimated distance and the observation reported by a sensor were to be provided in a stochastic form, then the model could be appropriately modified to incorporate these densities through integration. Assuming that in the general case a computer implementation would require numerical evaluation of the model, and thus the densities are likely to be represented in point-value form, real-time computation of $f_{I\,|\,D,V}$ for such a model is not likely to be feasible for a large number of points in the range of *i*. This limitation is likely to impose a serious restriction on both the amount of score considered at any given time and the resolution of the estimated density. It may reduce the accuracy of the final estimated density as well, since numerical integration is involved.

In order to increase the likelihood that this model can adequately accommodate many arbitrary density functions represented with sufficient resolution in point-value form, it will be assumed that the estimated distance, *d*, and the observation reported by the sensor, *v*, will take on a single value during each evaluation of the final conditional density, $f_{I\,|\,D,V}$. These values will be indicated by d_1 and v_1, respectively. To explicitly notate this fact, assignment notation will be used within equation 2.13 as follows:

$$f_{I\,|\,D,V}\,(i|\; d=\mathrm{d}_1,\; v=\mathrm{v}_1) = \frac{f_{V\,|\,I}\,(v=\mathrm{v}_1|\; i)\cdot\int_{j=0}^{\|Score\|} f_{I\,|\,D,J}\,(i|\; d=\mathrm{d}_1,\; j)\cdot f_{Prior}(j)\,\partial j}{\int_{i=0}^{\|Score\|} f_{V\,|\,I}\,(v=\mathrm{v}_1|\; i)\;\cdot\;\left[\int_{j=0}^{\|Score\|} f_{I\,|\,D,J}\,(i|\; d=\mathrm{d}_1,\; j)\cdot f_{Prior}(j)\,\partial j\right]\partial i} \qquad \textbf{[2.14]}$$

The numerical representation of the density functions containing these variables can thus be reduced in dimensionality for purposes of computation. Specifically, only the function $f_{I\,|\,D,J}$ will require a representation with more than one dimension. This simplification should significantly reduce the computation required both to evaluate $f_{I\,|\,D,V}$ and to generate the component density functions appearing in the right-hand side of equation 2.14.

2.8 A Convolution Assumption

The function $f_{I\,|\,D,J}$ in equation 2.14 remains a function of two variables, *i* and *j*. Consequently, if this function is represented in a point-value form, evaluating the integral over *j* for multiple values of *i* will require a number of arithmetic operations equal to the product of the number of points in the ranges of *i* and *j*. In addition, it is certainly not acceptable to assume that this function will be identical across all values of *j* for a given value of *d*. This means that either a full two-dimensional matrix representing this function for a given *d* will have to be generated during each evaluation of $f_{I\,|\,D,V}$, or a three-dimensional matrix specifying the entire function over the range of *d* will have to be generated in advance. Such an explicit representation of the function over an entire score is likely to be problematic.

In order to alleviate these implementation problems, we will make an additional assumption regarding the function $f_{I\,|\,D,J}$. Rather than considering the value of this function to depend on the source and destination positions individually, we will assume that the function can be adequately approximated by another function that depends on only the *actual* distance between the source and destination positions, $i - j$. Equation 2.15 defines this equivalence:

$$f_{I\,|\,D,J}\,(i|\,d,j) \;=\; f_{I-J\,|\,D}\,(i-j|\,d) \qquad \text{[2.15]}$$

Although at first this may seem like an unfounded assumption, it corresponds to a simple and intuitive explanation. If one knew the exact distance traversed in the score, one could simply shift forward the previous score position density by that amount. Since the distance is not known exactly, the previous score position density should be shifted and "smeared out" to reflect the uncertainty. As long as the amount of the "shift and smearing" is independent of the previous position, the equivalence in Equation 2.15 holds. This equivalence means that for a given value of *d* (the estimated distance), the value of the function $f_{I\,|\,D,J}$ is identical whenever the actual distance between the source and destination positions is identical. It does not matter if the performer is near the beginning of the score or somewhere in the middle; the only relevant information is the actual score distance traversed since the last update. This idea is depicted by the graphs in Figure 2-3. Note that the function $f_{I-J\,|\,D}$ is a conditional density of actual distance conditioned on estimated distance.

The advantage of this assumption is that equation 2.15 enables us to rewrite equation 2.14 as follows:

$$f_{I\,|\,D,V}\,(i|\,d=\mathrm{d}_1,\;v=\mathrm{v}_1) \;=\; \frac{f_{V\,|\,I}\,(v=\mathrm{v}_1|\,i)\cdot\int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\,d=\mathrm{d}_1)\cdot f_{Prior}(j)\,\partial j}{\int_{i=0}^{\|Score\|} f_{V\,|\,I}\,(v=\mathrm{v}_1|\,i)\;\cdot\;\left[\int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\,d=\mathrm{d}_1)\cdot f_{Prior}(j)\,\partial j\right]\partial i} \qquad \text{[2.16]}$$

so that the integral over j becomes a *convolution integral*. This substitution is important because there exist well-known algorithms for efficiently approximating convolution integrals based on a single representation of the shifted function (such as the one depicted in Figure 2-3). One such algorithm is the fast Fourier transform, or FFT. The application of this technique is discussed at length in the subsequent section of this chapter.

One possible concern about the assumption that the density function $f_{I\,|\,D,J}$ is equal to $f_{I-J\,|\,D}$ relates to the finite length of the score. If the performer is assumed always to be within the score, then a distribution over the actual distance moved must certainly change with the performer's score position since the range of possible values for actual distance will change. When performers are near the beginning of the score, it is theoretically possible for them to move a much farther distance in the forward direction than when they are near the end of the score. In reality, however, the tracking system will need to update its estimate of the performer's location so frequently that the overwhelming majority of the probability density will be located in an extremely short segment of the range of $i - j$, where the actual distance is quite small. The density will taper off so quickly that significant effects on the model could only occur when the performer is very near to the beginning or the end of the score. In addition, a further important property of this density (to be discussed in a subsequent chapter) will eliminate possible problems near the beginning of the score. The possibility of encountering problems very near the end of the score will prove to be of no concern.

This simplification completes the development of a continuous stochastic model for score following. A graphical example of applying the model is given in Figure 2-4. Throughout subsequent

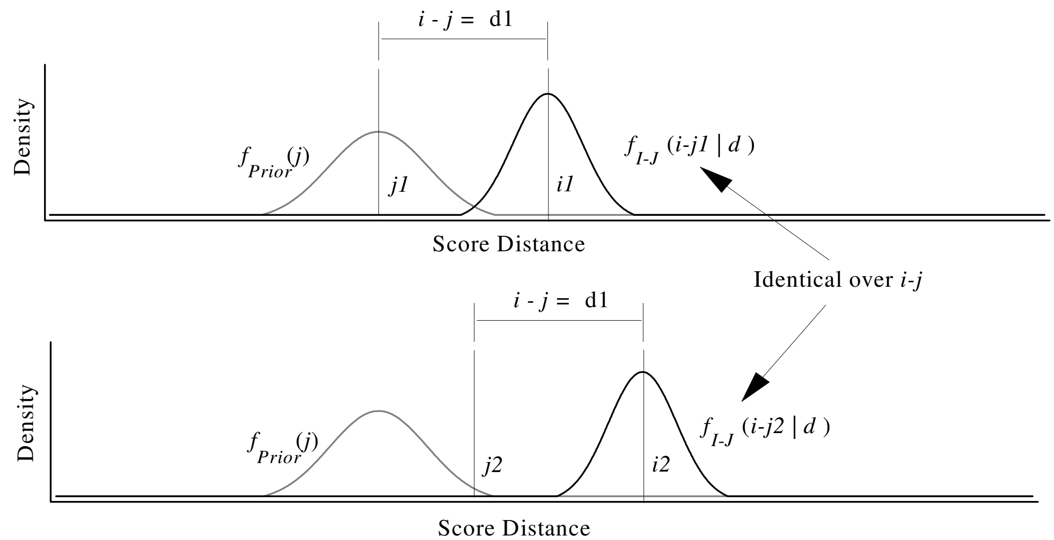


Figure 2-3. Depiction of the convolution assumption. The function f_{I-J} is identical over the actual distance, $i - j$, regardless of the source and destination positions.

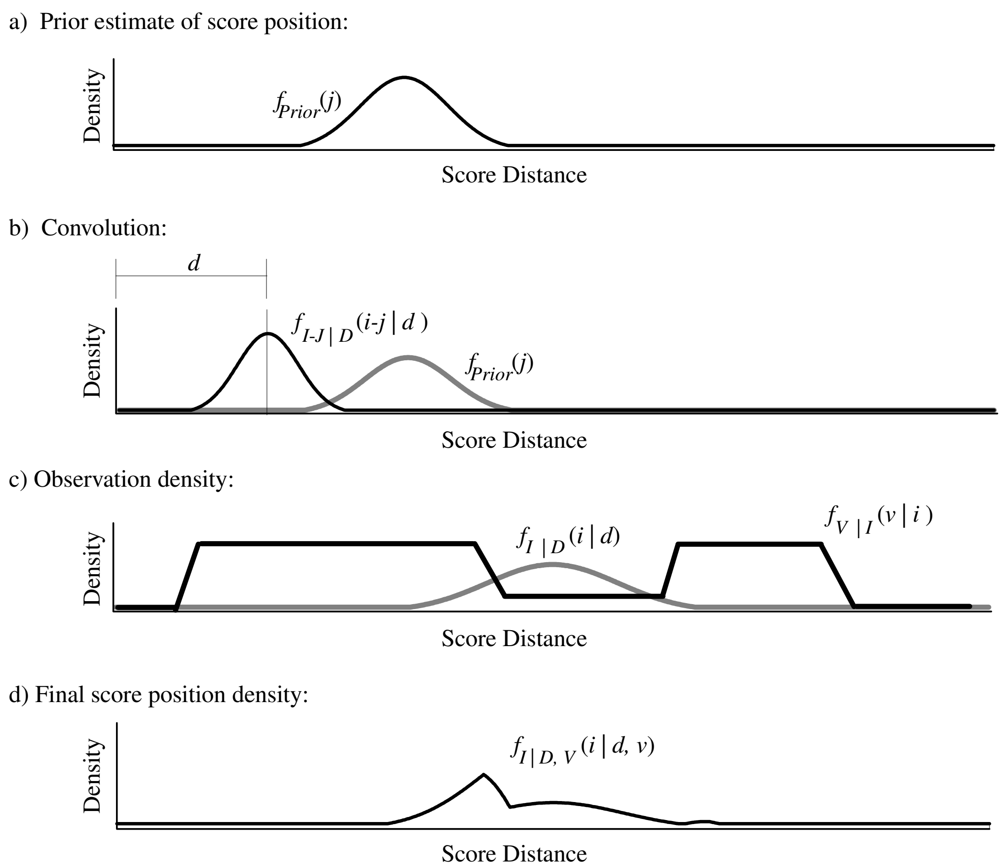


Figure 2-4. Depiction of a single application of the final stochastic score-following model. Graph a) shows the density over the performer's previous location. Graph b) shows the densities participating in the convolution. Graph c) shows the results of convolution and the observation density. Graph d) depicts the final score position density after completing a single application of the model.

discussion, the model will not be altered from its appearance in equation 2.16. The remaining sections of this chapter will describe a discrete approximation to this model that can be efficiently implemented on personal computers available today. To enhance the reader's intuition about the continuous model, some graphical examples are provided. Finally, we consider possible errors introduced as a result of both the assumptions previously made and some additional approximations. These approximations are required in order to actually implement the model and define the density functions.

2.9 A Discrete Approximation and Implementation

The general model presented in equation 2.6 gives rise to the model of equation 2.16 under the assumptions that the observations are independent of the source position and estimated distance, the observation and estimated distance are reported as single values, and the function $f_{I|D}$ can be well approximated by a convolution integral. In this section, a discrete approximation to the continuous model of equation 2.16 is presented along with graphical examples of this model in operation. It is shown that the discrete approximation permits a direct computer implementation of the model. In addition, a slight modification to this implementation, based upon the FFT, is shown to enhance the computational efficiency of applying the model.

It should be pointed out that while the approach taken here is to apply a discrete approximation in order to implement the model, an alternative would be to try to define all density functions in equation 2.16 using parameterized density functions. If appropriate families of functions could adequately represent these densities, then it might be possible to implement the model through calculations over the parameters alone. Unfortunately, such an approach would require that either the selected families of functions always yield members of those same families under the operations of the model, or that it is feasible to apply some manageable series approximation to the result functions. For the general case of score following, such a method is unlikely to be any less cumbersome than the discrete approach. The difficulty is due mainly to the nature of the function $f_{V|I}$ that for musical scores does not readily submit to representation by a simple parameterized density. As shown in previous examples, this function often contains multiple sharp transitions between plateaus in the density. The shape of this function has a significant impact on the final score position estimate produced by the model and should be approximated accurately.

The discrete approximation to the continuous model is based upon a point-value representation of all density functions. As depicted in Figure 2-5, it is assumed that the functions are sampled at an appropriate, constant interval over the ranges of the variables they contain. The functions $f_{I|D,V}$, f_{Prior}, $f_{I-J|D}$, and $f_{V|I}$ will be sampled at identical rates along their score position related dimensions, with a sample interval indicated by Δs. The variable S represents the number of sample points in the score. Since the range of the sampling for these functions will depend upon the length of the score, the minimum number of samples for all of these functions will be indicated by S. Sampling along other dimensions of

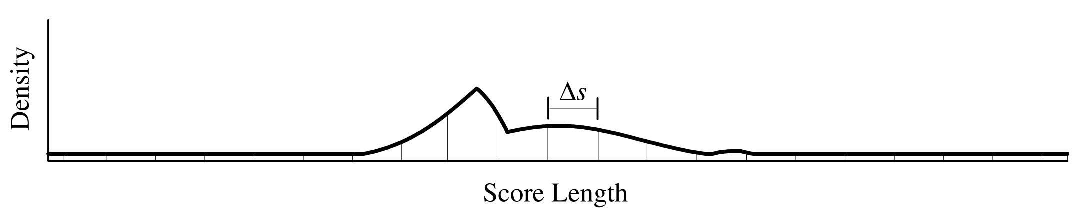


Figure 2-5. Sampling a density function that characterizes the score position of a performer. The sample interval is indicated by Δs.

these functions is not relevant to this section, but is discussed in subsequent chapters where estimates of the functions are provided.

To assist in the development and discussion of the discrete model, we decompose the final continuous model presented in equation 2.16 as follows:

$$f_{I\,|\,D}\,(i|\;d = \mathrm{d}_1) \;=\; \int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\;d=\mathrm{d}_1)\cdot f_{Prior}(j)\,\partial j \qquad \texttt{[2.17]}$$

$$f_{I\,|\,D,V}\,(i|\;d=\mathrm{d}_1,\;v=\mathrm{v}_1) \;=\; \frac{f_{V\,|\,I}\,(v=\mathrm{v}_1|\;i)\;\cdot\;f_{I\,|\,D}\,(i|\;d=\mathrm{d}_1)}{\int_{k=0}^{\|Score\|} f_{V\,|\,I}\,(v=\mathrm{v}_1|\;k)\;\cdot\;f_{I\,|\,D}\,(k|\;d=\mathrm{d}_1)\,\partial k}$$

reflecting the fact that during the actual computation, the integral over *j* is computed only once.

Now the discrete model will approximate all integrals appearing in the continuous model. This is done by summing estimated area under the curve within regions of size Δs centered around each sample. Area approximation using rectangles is depicted graphically in Figure 2-6. Sigma notation will be used in place of the integral signs:

$$\tilde{f}_{I\,|\,D}\,(i\cdot\Delta s|\;d=\mathrm{d}_1) \;=\; \sum_{j=0}^{S} f_{I-J\,|\,D}\,(i\cdot\Delta s - j\cdot\Delta s|\;d=\mathrm{d}_1)\cdot\tilde{f}_{Prior}\,(j\cdot\Delta s)\cdot\Delta s \qquad \texttt{[2.18]}$$

$$\tilde{f}_{I\,|\,D,V}\,(i\cdot\Delta s|\;d=\mathrm{d}_1,\;v=\mathrm{v}_1) \;=\; \frac{f_{V\,|\,I}\,(v=\mathrm{v}_1|\;i\cdot\Delta s)\;\cdot\;\tilde{f}_{I\,|\,D}\,(i\cdot\Delta s|\;d=\mathrm{d}_1)}{\sum_{k=0}^{S} f_{V\,|\,I}\,(v=\mathrm{v}_1|\;k\;\cdot\;\Delta s)\;\cdot\;\tilde{f}_{I\,|\,D}\,(k\;\cdot\;\Delta s|\;d=\mathrm{d}_1)\;\cdot\;\Delta s}$$

Note that while the continuous density functions still appear in equation 2.18, the score position variables *i*, *j*, and *k* always assume integral values over the defined range of samples. They will always be multiplied by the sample interval to indicate that only the sampled points of the density function are

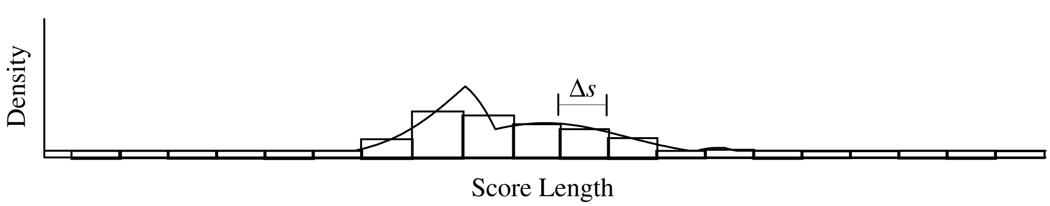


Figure 2-6. Rectangular integration of a density function. The area under the function is approximated by summing the area of contiguous rectangles centered at the sample points.

relevant to the computation. Density functions that are affected by the discrete calculation of the integrals, and therefore only *approximate* the equivalent function in the continuous model, will be marked with a tilde. The next section in this chapter provides a more complete discussion of how these functions vary from their counterparts in the continuous model.

In a final implementation of the model, several factors will influence the selection of the sample interval, Δs. First, the sampling interval must be small enough to give good resolution for purposes of estimating score position. If the interval is too large, interpolation between sample points may be needed in order to obtain sufficiently precise estimates of the performer's location. Depending upon the general shape of the density function over the score, some method of interpolation may or may not be an acceptable alternative to increasing the number of samples. Second, the accuracy of the model may degrade to an unacceptable level if the sample interval is too large. Reasonable approximation of the integrals in the model will require fairly fine sampling. Approximation will be discussed more formally in the next section. Finally, if the sample interval is too small, the computation time required for each update of the model may exceed the requirements for real-time performance, causing the system to fall so far behind in its "data processing" that it will never recover. Interestingly, as is discussed in the next section, an extremely small sample interval may also increase the inaccuracy of numeric estimates of the integrals contained in the model.

Figure 2-7 shows a flow chart corresponding to the implementation of the discrete model in equation 2.18, along with the computational complexity of each step. Note that allowance is made for runtime generation of both the distance density, $f_{I-J|D}$, and the observation density, $f_{V|I}$. The bottleneck in

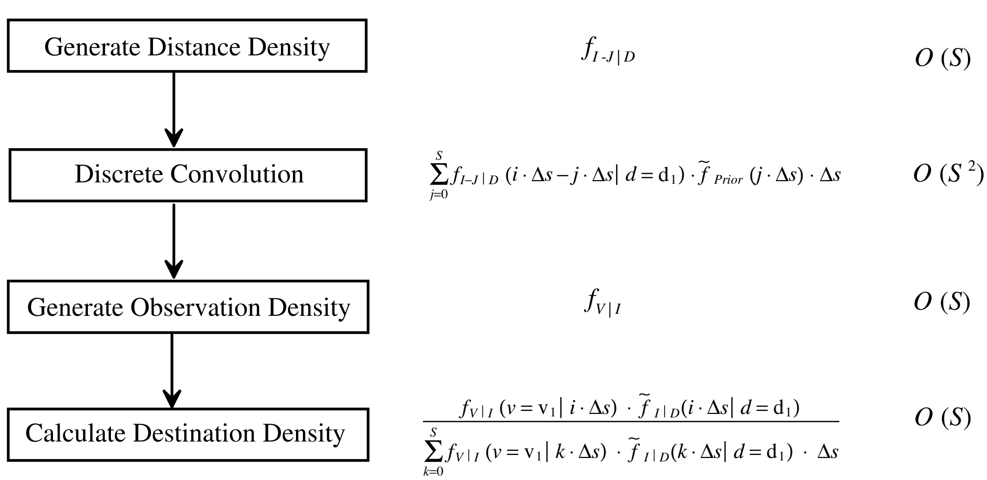


Figure 2-7. Flowchart for direct implementation of the discrete model.

this process is the straightforward evaluation of the integral over the source position, j. The value of the integral must be calculated for each possible destination position, i. In total, direct calculation of the discrete model is a quadratic-time (in S) operation. However, because we have allowed this calculation to be a convolution operation, where for each successive destination position the function $f_{I-J\,|\,D}$ is simply shifted one sample to the right, we can apply the fast Fourier transform to obtain a significantly more efficient implementation.

Figure 2-8 shows a flowchart for an implementation using the FFT. The Fourier transform of a function has an interesting property with respect to convolution involving that function. The Fourier transform of a convolution of two functions is equivalent to the product of the Fourier transforms of those functions. As shown in the flowchart, the model of equation 2.18 can be implemented by using the FFT

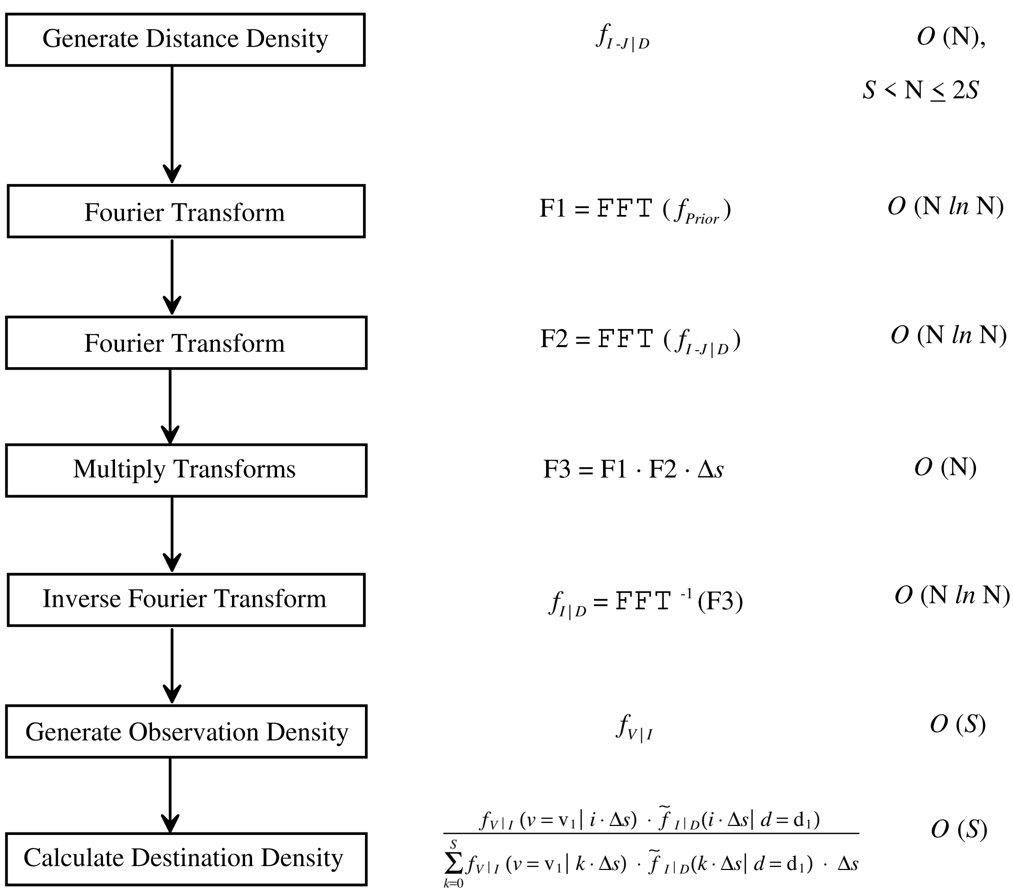


Figure 2-8. Flowchart for implementing the discrete model by using the fast Fourier transform. The Fourier transforms of f_{Prior} and $f_{I-J\,|\,D}$ are computed and multiplied. This product is multiplied by the sample interval. The inverse Fourier transform is computed to return $f_{I\,|\,D}$. Note that the number of samples is increased to N to avoid unwanted effects of the cyclic convolution.

to compute the discrete Fourier transforms of f_{Prior} and $f_{I-J\,|\,D}$, multiplying these two transforms, and then computing the inverse Fourier transform of the product by using the inverse FFT. Because the FFT can be implemented to execute with complexity nearly linear in the number of samples, this method of implementing the model is likely to be more efficient for even moderate numbers of samples ($N \geq 100$).

The Fourier transform is defined in both a continuous and discrete form. The convolution property just stated applies in both cases. The discrete version of convolution using the Fourier transform can be used to calculate the summation in equation 2.18 without multiplying by the sample interval. However, this multiplication can be extracted outside the summation and applied to the calculated sum. This property enables both implementations to yield identical results.

Discrete convolution via the Fourier transform differs from the direct summation of equation 2.18 in that $f_{I-J\,|\,D}$ is not only flipped and shifted as j increases, but is also in effect "wrapped around". This operation is referred to as *cyclic convolution* and is depicted in Figure 2-9. Because of this wrapping, calculation of the convolution using the FFT requires that the sampled representation of f_{Prior} be extended with zero-valued samples. This "packing" prevents the wrapped samples of the function $f_{I-J\,|\,D}$ (the points beyond the boundaries of the score) from influencing the discrete approximation of the integral. Packing of the function is depicted in Figure 2-10.

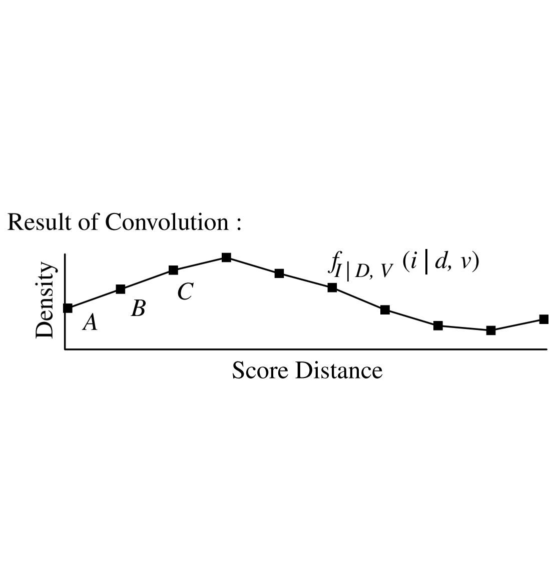


Figure 2-9. Cyclic convolution as results from executing discrete convolution via the Fourier transform. The graph on the left depicts the result of convolution. The graphs on the right depict the functions participating in the convolution. The function containing *x1* is shifted to the right by one sample for each point calculated by the convolution.

It is also evident from the first summation in equation 2.18 that if both the source position, j, and the destination position, i, range from zero to S; then the actual distance, $i - j$, will range in value from $-S \cdot \Delta s$ to $S \cdot \Delta s$. In the worst case, the sampled representation of the density $f_{I-J\,|\,D}$ must include all samples in this range, or a total of $2S$ samples. For this reason, the value of N in the flow chart using the Fourier transform can be as large as $2S$. In reality, if the density function $f_{I-J\,|\,D}$ quickly tapers off in both directions and either settles to a constant or falls below an acceptable minimum value for implementation purposes, the number of samples required for $f_{I-J\,|\,D}$ will be the score size S plus the number of samples until the minimum is reached in the forward direction. These extra samples will provide sufficient room for the samples with a value above the minimum to wrap around and not affect the values of the convolution corresponding to actual score positions. This situation is depicted in Figure 2-11.

One final point about an implementation using the Fourier transform is that the FFT algorithm can be implemented most easily if the number of samples for each function is a power of two. This limitation will not pose a problem for score following since we are most concerned about the sample interval and in general will not be able to process an entire score in real time. This latter restriction will require some sort of windowing over the score, the nature of which is subsequently discussed. The size of the window will of course be limited by the amount of computation that can be done in real time. It is fine if the limit on the number of samples happens to be a power of two as long as a useful window size can be processed on each iteration. No further details regarding the Fourier transform or its implementation will be provided here. Bracewell (1986) and Brigham (1974) provide excellent discussions of both the continuous and the discrete versions. Details of efficiently implementing the FFT are given by Mitra and Kaiser (1993) and Oppenheim and Schafer (1975).

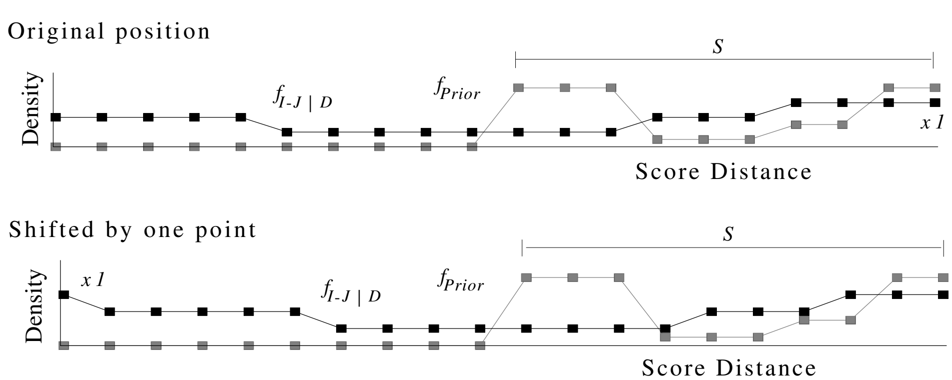


Figure 2-10. Extending f_{Prior} with zero-valued samples. During the cyclic convolution, the points that "wrap around" will not affect the calculation since they will be multiplied by zero.

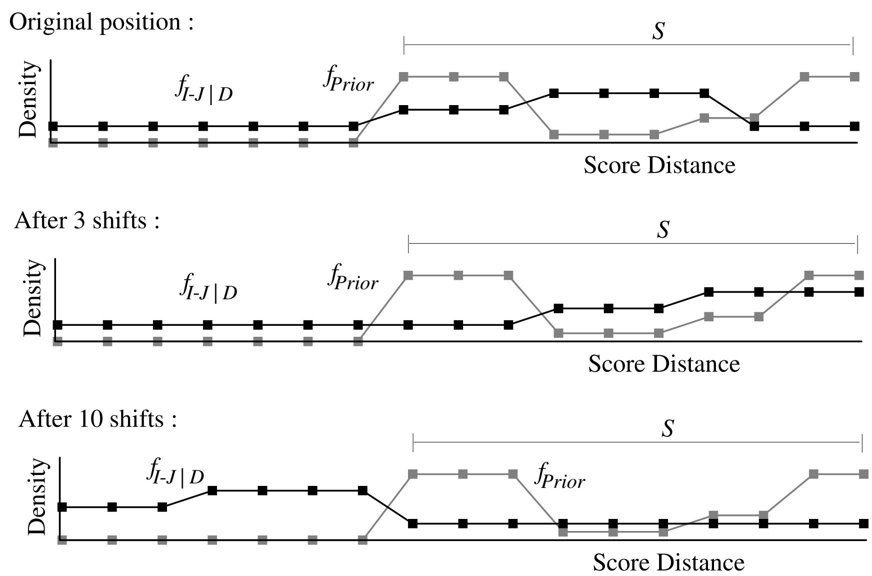


Figure 2-11. The uppermost graph shows f_{Prior} extended by only seven zeros, the number of points within S before $f_{I-J\,|\,D}$ reaches its minimum value. The middle graph depicts the functions after three shifts during the convolution operation. The lower graph shows the functions after ten shifts (the total number of meaningful points to be calculated by the convolution) demonstrating that the wrapping will not impact the results of interest.

It is important to note that rectangular approximation of area under the curve may cause unacceptable inaccuracies when using samples that fall on the endpoints. Because the very first and the very last points in the score will have a density of zero, estimates of the area under the curve in regions centered around these points will always be zero. In cases where this estimate is unacceptably low, two possible solutions exist. The first is to simply use the value of the density at the point halfway between the boundary point and the next (or previous) sample, divided by two. This approach will give an approximation for the area in the boundary region that spans only half a sample interval. The second solution is to offset the sampling by half of a sample interval at the beginning of the score, and to either dismiss the error incurred by the score region associated with the last sample, or to adjust the density for this sample as per the first solution. It should be noted that these proposed adjustments to the density function, f_{Prior}, can be applied directly to the sampled function before computing its Fourier transform. Calculating convolution using the modified transform yields an adjusted estimate of area identical to that produced using direct calculation of the convolution.

An improved estimate of the integral can often be obtained by applying Simpson's rule for integral approximation. Rather than calculating area of rectangles centered around each sample, Simpson's rule approximates the integral by summing the areas under parabolas fit to sets of three sampled points. Successive sets share a common point so that the density curve is approximated by contiguous partial parabolas. The Simpson approximation for the area under each set of three points is:

$$A = \frac{\Delta s}{3} \cdot [f(x_1) + 4f(x_2) + f(x_3)] \qquad [2.19]$$

where f indicates the function whose integral is to be approximated and the numbered variables x indicate the points in the first set. By summing these approximations for all sets of points, the following expression for estimating the integral is derived:

$$A = \frac{\Delta s}{3} \cdot [f(x_1) + 4f(x_2) + 2f(x_3) + 4f(x_4) + \cdots + 4f(x_{S-1}) + f(x_S)] \qquad [2.20]$$

Note that the number of samples, S, must be odd. The convolution operation that is implemented via the FFT can be easily modified to approximate integrals using this formula. Since the point x_1 will always be the first point in the score, regardless of the shifting that occurs during convolution, f_{Prior} can be multiplied by the coefficients in equation 2.20 before its transform is computed. The result of convolution can then be multiplied by one-third the sample interval to complete the estimate. Note that this extra computation does not significantly impact the previous complexity analysis for the implementation using the FFT, since this new computation only requires S additional multiplications prior to the convolution. As will be shown in the subsequent section, this additional computation is likely to be trivial compared to the improved integral approximation that often results from this formula.

A final comment about implementation of the model is that real-time calculation of the model over an entire score is unlikely to be possible, even using the FFT. To allow for real-time application of the model, an implementation can apply a windowing technique similar to that described by Dannenberg (1984). This technique was applied to the dynamic programming algorithm for score following. Instead of calculating the density $f_{I|D,V}$ over the entire score on each iteration, the implementation will calculate the density within a reduced window of the score. Ideally, the window is small enough to permit real-time computation of the model but large enough to encompass those score locations that have significant likelihood of representing the performer's current position. On each iteration, the density $f_{I|D,V}$ will be calculated over a region of score that is shifted slightly from the region calculated for the density f_{Prior}. This new region will be centered around the most likely source position of the performer, as decided upon by the accompaniment system. This point is likely to correspond closely with the center of a localized region of high density in f_{Prior}. Shifting of the window is depicted in Figure 2-12.

Implementing the shifting of the window is straightforward since this only requires shifting one of the density functions in the appropriate direction prior to convolution. This approach is depicted in Figure 2-13. The technique enables estimation of the model at points outside of the previous window using only points within that window. Providing the window size is sufficiently large so that points beyond the window are not likely to contribute significantly to the calculation of the first integral in

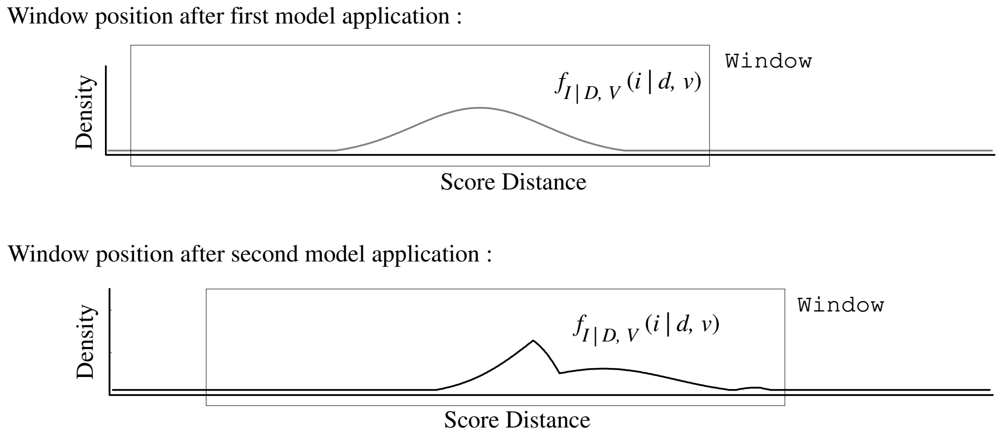


Figure 2-12. Depiction of windowing the score. The upper graph shows a window of the score containing the most likely locations of the performer. During subsequent application of the model, the window is re-centered. The lower graph shows that if the window is sufficiently large, the most likely locations of the performer will remain within the window.

equation 2.18, the results of the windowed implementation should closely approximate those of an implementation that calculates the density over the entire score.

The only real difficulty with windowing is that if the performer should ever jump outside of the current window, it is likely that the tracking system will remain lost until the performer moves back inside the window. An *ad hoc* solution to this problem is to use multiple windows at places where the performer might jump to a new location (Dannenberg and Mukaino 1988). A more general and formally described recovery mechanism is desirable, but this problem is not addressed as part of this work.

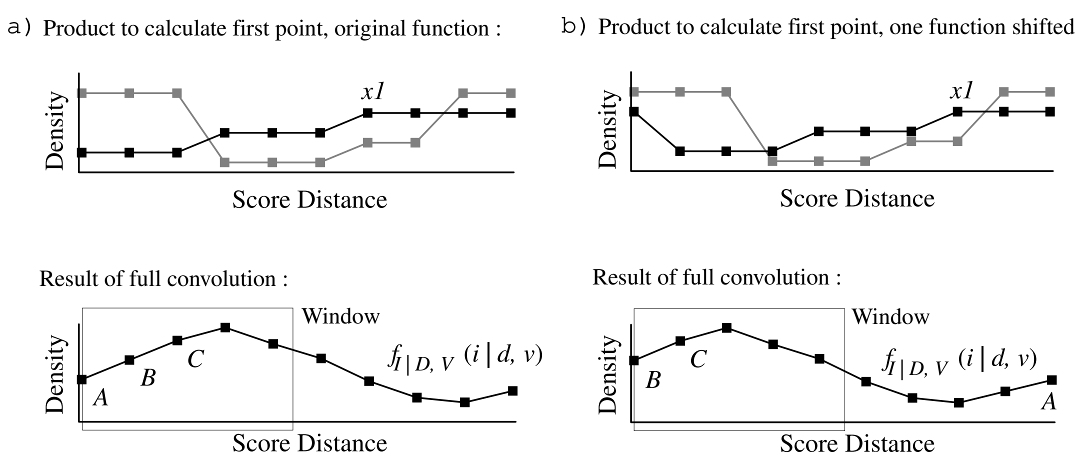


Figure 2-13. Effecting the window movement by shifting one function prior to convolution. The pair of graphs a) show the results of convolution from an original positioning of the functions. The graphs in b) show the results after one function has been shifted to the right by one point *prior* to calculating the convolution. As shown in b), the results inside the window will be shifted one point to the left.

2.10 A General Error Analysis of the Models

The performance of a score-following system strongly depends upon how accurately and precisely an implemented model characterizes the real world. The preceding sections have described and related three versions of a model for stochastic score following—a continuous model, a discrete model, and a model of implementation. The continuous model is appealing because it offers the most potential both to reliably approximate the real world and to present a constructivist view of the score-following process that is most concerned with the domain and least concerned with a computer implementation. It is the most appropriate model for discussing the interaction of density functions and how they should be defined. The discrete model provides a description of score following appropriate for implementation on a theoretical computer unhindered by resource constraints. It highlights the differences between a continuous view of the world and one that is theoretically appropriate for implementation. The model of implementation provides a fully pragmatic view, modified from that of the discrete model to reflect the resource limitations of the available computer platform.

In transitioning between these three models, several possible sources of error have been introduced. These constitute adjustments to the view of the problem. They have the potential to influence both the precision and the accuracy of a model relative to the actual behavior of performers and sensors. The possible sources of error introduced by transitions between models are depicted in Figure 2-14. Complete specification of the continuous model allows for three general sources of error. First, the model will be less accurate if some relevant information is not defined as a variable in the model. If gender, voice part, or experience of the singer were to affect the viability of the convolution assumption, for example, then incorporating this information in the general model for score following might influence our decision to make that assumption. Second, errors can arise from assumptions needed to justify a formal

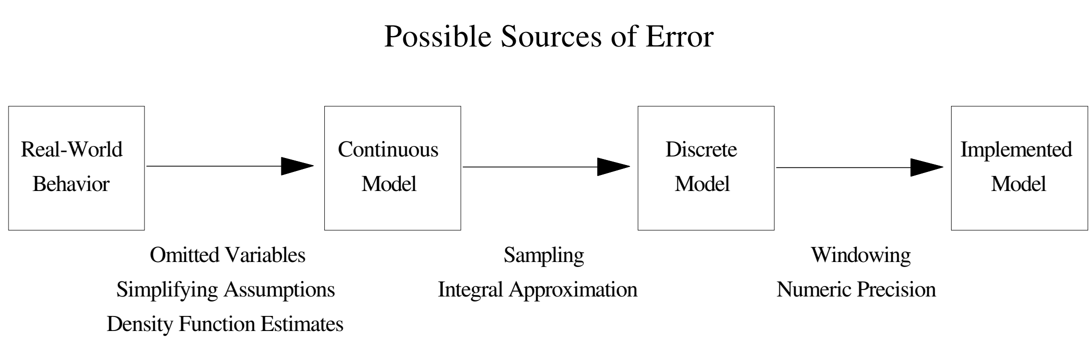


Figure 2-14. Possible sources of error introduced by transitioning between the various models.

description of the model that is manageable both conceptually and in practice. The latter type of manageability simply means that the model can be defined at a level of precision and completeness that enables implementation on a computer. Finally, the way in which the model's density functions are defined can significantly impact how accurately the performer's position is predicted.

The sources of error within the continuous model have been discussed throughout the development of the model. In general, the extent and magnitude of these errors can only be determined empirically. It is important to consider them when attempting to estimate the density functions in the model. These functions include the densities describing both the actual distance moved by the performer and the observations reported by the sensors. In subsequent chapters, consideration is given to whether these densities provide support for or against the assumptions used to convert the general model into the final continuous model. The evaluation includes an assessment of tracking accuracy when applying the score-following model to actually accompany singers. This analysis provides useful insight into which of the possible sources of error in the continuous model significantly impacts position prediction.

Converting the continuous model into a discrete model introduces two possible sources of error. The first error results from the sampling of the continuous functions. If updating the model sometimes requires the value of a function at a point not explicitly sampled, then an approximation to the actual value must be used. This problem does not arise for the density function describing actual distance because the discrete model requires only values for distances between all sampled source and destination points. Such a set of distances is well-defined and finite, and in general can be easily accommodated. Depending upon the type of sensor observations reported, this also may be true of the density function describing observations. In the case of the final density function defined over the destination position of the performer, some error would be introduced if values of the function at nonsampled score positions must be approximated. Such approximation would be necessary, in order to give a useful stochastic estimate of the performer's location, if the density function contained sharp peaks or troughs between samples. If the sampling is fine enough to guarantee that the value of the function between samples is not significantly different from the value at the bounding samples, then sampling is not a serious problem. This general idea is depicted in Figure 2-15.

Another error that results indirectly from the sampling is the error in the approximation of the integrals. All of the density functions are involved in integration, all of these integrals are calculated over a variable of score position. In addition, determining the probability that a performer is within a given region of the score requires integration of the destination position density after each update of the model. While numerical bounds can be placed on the error of both rectangular integration as applied in the discrete model and Simpson's rule, these estimates are based on specific information about the functions to

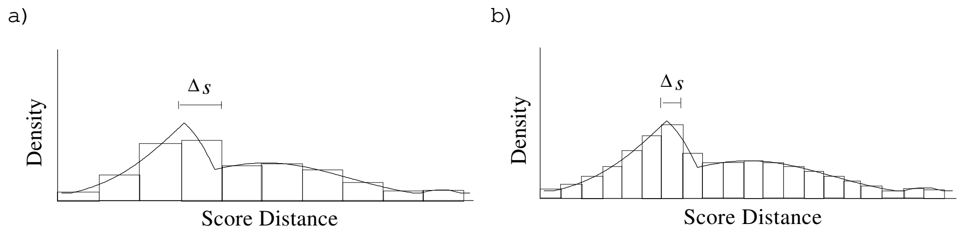


Figure 2-15. Sampling the same function using two different sample intervals. Graph a) depicts sampling with a large interval while graph b) depicts sampling with a smaller interval. Graph a) shows a larger discrepancy between the density at the center of the rectangles and the density at the edges.

be integrated—namely properties of their higher-order derivatives. This information has not yet been specified for the functions in the score-following model. If f is the integrand and [a,b] the range of integration, then the error of the integral estimate calculated via rectangular integration as previously described (often referred to as *midpoint approximation*) is given by:

$$\frac{(b-a)^3}{24(S-1)^2} \cdot f''(x), \quad a < x < b \tag{2.21}$$

The error of the integral estimate calculated using Simpson's rule is given by:

$$\frac{(b-a)^5}{180(S-1)^4} \cdot f^{(4)}(x), \quad a < x < b \tag{2.22}$$

These error expressions hold providing the respective derivatives exist. Note that even in the worst case, the error from the rectangular approximation will be proportionate to $(S-1)^{-2}$ and the error from integration via Simpson's rule will be proportionate to $(S-1)^{-4}$. Thus, increasing the number of samples used to approximate the density functions will thus reduce the error at a fairly rapid rate. In the case of Simpson's rule, the error will decrease by more than an order of magnitude every time the sample rate is doubled. Consequently, if an approximation using 5 samples (2 parabolas) provides at least 1 decimal place of accuracy, then we should expect about 4 decimal places of accuracy when using at least 33 samples (16 parabolas).

However, there is limited value in increasing the sample rate. This limitation is due to the roundoff error that is introduced when converting the discrete model into the model of implementation. Roundoff error is a result of using a finite computer representation for an infinite set of values. When a large number of arithmetic operations contributes to determining a single value, roundoff error can be substantial. Davis and Rabinowitz (1967) provide a theoretical bound for roundoff error of integral approximation techniques of the kind applied here. While actual numerical values given by this bound are likely to be overly conservative, the analysis does indicate that the error grows no worse than linearly with the number of samples. However at some point, as the number of samples is increased, the magnitude of the roundoff error will eventually surpass the error due to the approximation. As a result, the entire error of the estimate will begin to increase. Oppenheim and Weinstein (1972) provide bounds for roundoff error of the discrete Fourier transform. This error is again trivial compared to the potential error of the integral approximations. Thus, for a representation of reasonable precision, it is unlikely that roundoff error will be a serious problem for either the direct implementation or the implementation using the FFT.

The final source of error, which arises from converting the discrete model to a model of implementation, is the error incurred by the windowing technique. As described in the previous section, this error results from approximating the integrals without including contributions from points outside the window. Providing that the window is both sufficiently large and positioned to encompass all regions of significant density, and the products inside the integrals rapidly taper off to extremely small values, the windowing should not contribute significantly to the integration error. The possibility that the performer will move outside of the current window is always a problem, but only to the extent that it is likely to happen. For a window spanning several seconds of score, this likelihood is insignificant.

While the errors introduced by the continuous model can only be dismissed through empirical investigation, there are two important facts about the general model and the use of its results that further reduce the significance of some of the remaining sources of error. First, both the continuous model and the discrete model are invariant under scaling of their component functions. In other words, if any of the density functions in the model is multiplied by a constant, this will not affect the value of the approximation of $f_{I|D,V}$. These constants will pass through the integrals and cancel in the final ratio of both equations 2.16 and 2.18. Second, the accompaniment system will not make decisions based upon the absolute value of points in the function $f_{I|D,V}$, but will only consider the relative value of such points for purposes of determining the regions more likely to encompass the location of the performer. This use of the function amounts to examining a ratio of two values of the function. Calculating such ratios will cancel all errors that amount to multiplying the correct function by a constant.

This latter calculation of a ratio will cancel any error that results from calculating the second summation in equation 2.18. Furthermore, all errors due to calculation of the second summation will not propagate through subsequent applications of the model when $f_{I|D,V}$ is substituted for f_{Prior}. This property is due to the model's invariance under scaling of the component functions. Only the error from calculating the convolution integral and roundoff error from operations outside the second summation can possibly affect the value of $f_{I|D,V}$ on each iteration. Only these errors can propagate through subsequent iterations.

To examine this more precisely, consider the following equation that presents the discrete model of equation 2.18 in a format appropriate for examining the propagation of error:

$$f_{I\,|\,D}\,(i\cdot\Delta s|\ d)\cdot\varepsilon(i)\ =\ \sum_{j=0}^{S} f_{I-J\,|\,D}\,((i-j)\cdot\Delta s|\ d)\cdot f_{Prior}\,(j\cdot\Delta s)\cdot\Delta s \qquad \textbf{[2.23]}$$

$$f_{I\,|\,D,V}\,(i\cdot\Delta s|\ d,\ v)\cdot\varepsilon(i)\cdot E\ =\ \frac{f_{V\,|\,I}\,(v|\ i\cdot\Delta s)\ \cdot\ f_{I\,|\,D}\,(i\cdot\Delta s|\ d)\ \cdot\ \varepsilon(i)}{\sum_{k=0}^{S} f_{V\,|\,I}\,(v|\ k\cdot\Delta s)\ \cdot\ f_{I\,|\,D}\,(k\cdot\Delta s|\ d)\ \cdot\ \varepsilon(i)\ \cdot\ \Delta s}$$

Note that for clarity, the use of equivalence notation inside the density functions has been dropped. All density functions represent the true continuous densities and not approximations. The function ε defines the error from approximating the convolution integral for the *i*'th sample of $f_{I|D}$. This error may include roundoff and windowing error as well as error due to approximating the integral. Note that this equation represents the error as a proportion of the true value and will always assume a positive value, hopefully very close to 1. The multiplier *E* represents the error introduced by approximating the integral in the denominator of the ratio. Note that the value of *E* is likely to be close to 1, and may be larger or smaller than 1 depending upon whether the summation underestimates or overestimates the integral. We temporarily ignore roundoff error that can result from the multiplication and division in the second part of equation 2.23.

Now the values of the function ε represent error from approximating the convolution integral. We will define two values ε_{max} and ε_{min} as follows:

$$\varepsilon_{\min}\quad\le\quad\varepsilon(i)\quad\le\quad\varepsilon_{\max} \qquad \textbf{[2.24]}$$

These values represent bounds on the error from approximating the convolution. Another value, *a*, will be defined as the weighted average of the error function, ε:

$$a\ =\ \frac{\sum_{i=0}^{S} f_{V\,|\,I}\,(v|\ i)\ \cdot\ f_{I\,|\,D}\,(i|\ d)\ \cdot\ \varepsilon(i)\ \cdot\ \Delta s}{\sum_{i=0}^{S} f_{V\,|\,I}\,(v|\ i)\ \cdot\ f_{I\,|\,D}\,(i|\ d)\ \cdot\ \Delta s} \qquad \textbf{[2.25]}$$

Note that the "weights" in the average correspond to the approximated area of $f_{I|D,\,V}$ using the true values of the sampled points in both the result of convolution and the observation density. Using this weighted average, we can account for the error in the model's estimate of $f_{I|D,V}$ as follows:

$$f_{I\,|\,D,V}\,(i \cdot \Delta s|\; d,\; v) \cdot \frac{\varepsilon(i)}{a} \cdot E' \;=\; \frac{f_{V\,|\,I}\,(v|\; i \cdot \Delta s) \;\cdot\; f_{I\,|\,D}\,(i \cdot \Delta s|\; d) \;\cdot\; \varepsilon(i)}{\sum_{k=0}^{S} f_{V\,|\,I}\,(v|\; k \cdot \Delta s) \;\cdot\; f_{I\,|\,D}\,(k \cdot \Delta s|\; d) \;\cdot\; \varepsilon(i) \;\cdot\; \Delta s} \qquad \textbf{[2.26]}$$

where E' accounts for the error of approximating the integral in the denominator using the true values of $f_{V|I}$ and $f_{I|D}$. Note that the ratio in the left-hand side of Equation 2.26 will either be less than 1 for some i and greater than 1 for other i, or equal to 1 for all i. We can now bound the error of the approximation as follows:

$$f_{I\,|\,D,V}(i \cdot \Delta s|\; d,\; v) \cdot \frac{\varepsilon_{\min}}{a} \cdot E' \;\le\; f_{I\,|\,D,V}(i \cdot \Delta s|\; d,\; v) \cdot \frac{\varepsilon(i)}{a} \cdot E' \;\le\; f_{I\,|\,D,V}(i \cdot \Delta s|\; d,\; v) \cdot \frac{\varepsilon_{\max}}{a} \cdot E' \qquad \textbf{[2.27]}$$

This provides both an upper and lower bound for the error on a single application of the model, assuming exact representations for all the required density functions, including f_{Prior}. Since these bounds are related to the errors from approximating the integrals, they can be made fairly tight for a single application of the model simply by reducing the sample interval, Δs. However, the lower bound may be less than $\varepsilon_{\min}$ or the upper bound greater than $\varepsilon_{\max}$ depending upon whether a is greater than 1 or less than 1, respectively. The error for the estimate of an individual point is not independent of the other points.

We can consider how these errors propagate over successive applications of the model. In a subsequent calculation of the model, the upper bound on the error of the approximation of $f_{I|D,V}$ will also be the upper bound on the error of the approximation to f_{Prior}. The error from calculating the convolution can thus be accounted for as follows:

$$f_{I\,|\,D}\,(i \cdot \Delta s|\; d) \cdot \varepsilon_n(i) \cdot \frac{E'}{a_{n-1}} \;=\; \sum_{j=0}^{S} f_{I-J\,|\,D}\,((i-j) \cdot \Delta s|\; d) \cdot f_{Prior}\,(j \cdot \Delta s) \cdot \frac{\varepsilon_{n-1}(j)}{a_{n-1}} \cdot E' \cdot \Delta s \qquad \textbf{[2.28]}$$

The subscript on the function ε indicates sequencing, so that ε_{n-1} is the error resulting from the previous calculation of the model. A subscript is also used on the weighted average, a. Note that the multipliers E' and a_{n-1}^{-1} are carried through the summation. They will cancel in the ratio calculated in the subsequent part of the model and will not affect the final result (other than by influencing the roundoff error). The only remaining error that affects the final result is that specified by the function ε_n.

We would like to place upper and lower bounds on the function ε_n. First, the expression defining the error can be simplified by removing the multipliers E' and a_{n-1}^{-1} from equation 2.28 and defining another weighted average, $\delta_n(i)$, as follows:

$$\delta_n(i) = \frac{\sum_{j=0}^{S} f_{I-J\,|\,D}\ ((i-j)\cdot\Delta s|\ d)\cdot f_{Prior}\ (j\cdot\Delta s)\cdot\varepsilon_{n-1}(j)\cdot\Delta s}{\sum_{j=0}^{S} f_{I-J\,|\,D}\ ((i-j)\cdot\Delta s|\ d)\cdot f_{Prior}\ (j\cdot\Delta s)\ \cdot\Delta s} \quad [2.29]$$

that will permit the following expression to account for the error in the convolution approximation:

$$f_{I\,|\,D}\ (i\cdot\Delta s|\ d)\cdot\varepsilon_n(i) = \delta_n(i)\cdot\sum_{j=0}^{S} f_{I-J\,|\,D}\ ((i-j)\cdot\Delta s|\ d)\cdot f_{Prior}(j\cdot\Delta s)\cdot\Delta s \quad [2.30]$$

By solving this equation for ε_n we obtain:

$$\varepsilon_n(i) = \delta_n(i)\cdot\frac{\sum_{j=0}^{S} f_{I-J\,|\,D}\ ((i-j)\cdot\Delta s|\ d)\ \cdot f_{Prior}\ (j\ \cdot\ \Delta s)\cdot\Delta s}{f_{I\,|\,D}\ (i\ \cdot\ \Delta s|\ d)} \quad [2.31]$$

We can further simplify this expression by observing that the ratio is just the error from approximating the convolution integral given an accurate estimate of f_{Prior}. If we allow $\varepsilon_{\max}$ to represent the maximum error incurred by approximating this integral for any n, and $\varepsilon_{\min}$ the minimum error, then both upper and lower bounds on ε_n can be given by:

$$\delta_n(i)\cdot\varepsilon_{\min} \quad\leq\quad \varepsilon_n(i) \quad\leq\quad \delta_n(i)\cdot\varepsilon_{\max} \quad [2.32]$$

The function ε_n will pass through the subsequent calculations of the model and will affect the result just as in equation 2.23. The right-hand side of equation 2.32 can thus be substituted in place of $\varepsilon_{\max}$ in equation 2.27 to place an upper bound on the error of the n'th application of the model. The left-hand side can be substituted for $\varepsilon_{\min}$ to give a lower bound on the error.

Now the propagation of error will depend on δ_n. As defined in equation 2.29, this is a weighted average of the $\varepsilon_{n-1}(j)$. If all values of ε_{n-1} are identical, then this weighted average will be constant for all i and will cancel in the ratio calculated in the subsequent part of the model. Otherwise, there is at least one value of ε_{n-1} below the average and at least one value of ε_{n-1} above the average, and δ_n will always be less than the maximum value of ε_{n-1}. This permits the following upper bound to be placed on δ_n:

$$\delta_n(i) \quad<\quad \max_j\,[\varepsilon_{n-1}(j)] \quad [2.33]$$

The exact value of δ_n will depend upon which values of ε_{n-1} are weighted most heavily in equation 2.29. In the absolute worst case, the highest values of ε_{n-1} will always be weighted most heavily for some point, and δ_n for that point will always be close to the maximum value of ε_{n-1}. In this case, the upper bound on the error over several applications of the model may be exponential in $\varepsilon_{\max}$. Even though $\varepsilon_{\max}$ may be extremely close to 1, after a certain number of iterations the possible error will become quite large. However, if sometimes $\delta_n << \varepsilon_{\max}^{n-1}$ for every point, then the growth of the maximum error may be substantially slowed. In particular, if the lower values of ε_{n-1} are sometimes weighted most heavily for every point, then δ_n may sometimes be less than 1 and sometimes above 1. If this is true, then the maximum propagated error may on average remain very close to $\varepsilon_{\max}$, possibly even within a constant power of it. An analogous line of reasoning can be given for the minimum error as well.

For n applications of the model, a definite upper bound on the relative error between any two points in $f_{I\,|\,D,V}$ can be given by the ratio of $\varepsilon_{\max}^{\ n}$ to $\varepsilon_{\min}^{\ n}$. A lower bound can likewise be given by the inverse of that ratio. If we assume that the maximum length of a song is around 8 minutes or 480 seconds, and the model will be applied as often as 10 times per second, the value of n will approach 5000. Assuming this estimate is more than conservative enough to account for the effects of the earlier dismissed rounding errors, and that $\varepsilon_{\min} = \varepsilon_{\max}^{-1}$, then $\varepsilon_{\max}$ must be less than 1.00007 in order to guarantee a maximum relative error of $u = 2$ when comparing points in the final estimate of $f_{I\,|\,D,V}$. The following equation describes this relationship between u and $\varepsilon_{\max}$:

$$u = \frac{\varepsilon_{\max}{}^{n}}{\varepsilon_{\min}{}^{n}} = \frac{\varepsilon_{\max}{}^{n}}{\left(\frac{1}{\varepsilon_{\max}}\right)^{n}} = \varepsilon_{\max}{}^{2n} \qquad \text{[2.34]}$$

For the example given, this means that the maximum error from approximating the convolution integral must be more than 4 orders of magnitude smaller than the true value of the point. The error bound on integration using Simpson's rule indicates we are likely to achieve this if the regions of significant density of the integrand encompass more than 33 samples. Still, in certain circumstances it may not be possible to reduce the error in the integral approximation by this much and also achieve real-time model calculation for a reasonably wide window over the score. However, empirical examination of the behavior of the functions in the convolution may support the belief that the propagated error is actually much less. Hopefully it does not compound at the maximum rate possible.

Although it would be inappropriate at this point to discuss details of implementation for specific computer platforms, it should be stressed that the combination of reducing the error by an order of magnitude through doubling the number of samples and the nearly linear complexity of the model's implementation is quite powerful. If the initially selected computer platform can achieve sufficient real-time integral approximation with 4 decimal place accuracy, then moving to a platform twice as fast could provide estimates with at least 5 decimal place accuracy. This improvement would be sufficient to satisfy the given (hopefully conservative) error bound for a value of $u = 2$ even when applying the model ten times per second for nearly one hour. Given a platform with sufficient numeric precision and speed to achieve integral estimates with 8 decimal place accuracy, the model could be continuously applied at the same rate for well over one month before exceeding the bound of $u = 2$.

While the errors introduced during a single calculation of the score-following model are likely to be negligible, the fact remains that the errors from calculating the convolution at least have the potential to propagate over successive evaluations. Consequently, one should be cautious about interpreting the results of successive applications of the model. The score position density should probably not be taken as a direct estimate of the actual continuous density without carefully considering details of the implementation. Caution should also be applied when comparing sample points whose values are well within an order of magnitude of one another.

Fortunately, for the purposes of score following, it will be important that the density function clearly distinguish the current location of the performer. Ideally, each estimate of f_{Prior} should have the overwhelming majority of the density over a very small region of the function. Likewise, the distance density used for convolution and the observation density should ideally have most of their density over small regions of the score. Such "sharpness" will enable these functions to provide the most help in discriminating between positions in the score. This idea is depicted in Figure 2-16 for the case of a fairly sharp observation density. Assuming the shape of the approximated f_{Prior} is relatively accurate, and the distance and observation densities are accurately modeled, selecting point B over point A as an estimate of

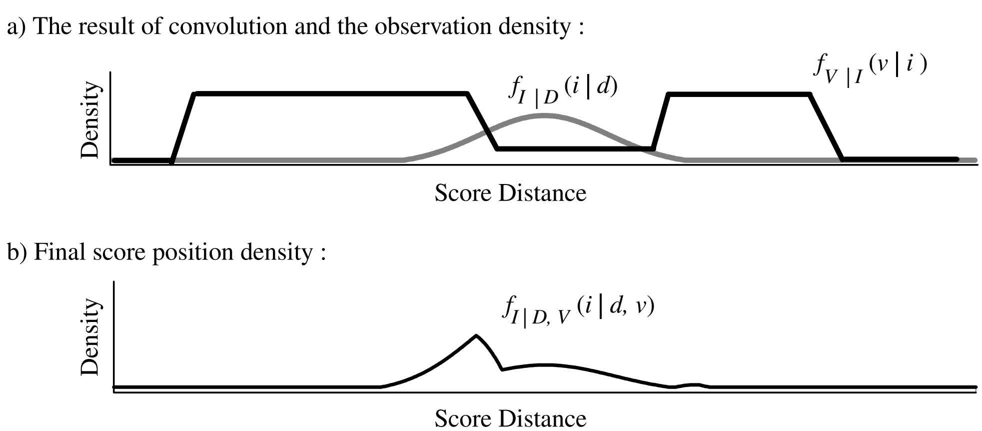


Figure 2-16. How sharp observation densities can dominate. Graph a) shows the result of convolution and the observation density. Graph b) shows the final result of the model, demonstrating how sharp observation densities can dominate distinction of event transitions.

the performer's position should not be suspicious given the large difference in magnitude between the density in regions around these points. However, selecting point B over point C might be viewed as more of a random choice. As long as subsequent application of the model keeps $f_{I\,|\,D,V}$ (and hence f_{Prior}) sharp, these essentially random choices will hopefully not affect the tracking system's performance. Whether or not this objective can be achieved will depend significantly upon the density functions that are used in the model.

2.11 The Process of Determining Distributions

The two pre-defined density functions within the stochastic score-following model, $f_{I-J\,|\,D}$ and $f_{V\,|\,I}$, have great influence over the accuracy of the score-following system. As can be seen from many of the graphic representations of updating the model, the shapes of these two functions can largely determine the areas of high probability in $f_{I\,|\,D,V}$. If these functions are very smooth and flat, they will provide little or no information about the performer's position. Convolution by a uniform density (a density function defined by a constant) will yield a uniform density, and thus the density over the performer's destination position will be dictated by the observation density. Likewise, if the observation density, $f_{V\,|\,I}$, is a uniform density, it will cancel out of the final ratio in the model; and the density over the performer's destination will be solely derived from the convolution.

Several subsequent chapters of this document pursue detailed definitions of the density functions for musical score following. Derivation of these functions is based in part on analysis of data collected from actual musical performances by vocalists singing with a live accompanist. The functions are defined within the context of tracking a trained vocalist singing pieces that have been rehearsed privately or with another accompanist. Selection of relevant parameters on which to base the density functions, such as the observations returned by sensors and the information in the score used to define the events, is based on how well the resulting functions help to discriminate score position. One important objective of testing the completed system is to demonstrate that a metric of the discriminatory ability of these functions can be identified. Such a metric will ideally permit prediction of average system performance when given the density functions used in the model and statistical descriptions of both the performer's location and the corresponding data to be processed by the system.

When defining the density functions, care is taken to insure as much as possible that the definitions do not violate any of the assumptions imposed by the various models. It is noted if the correctness or utility of any density function appears to have been compromised as a result of satisfying these assumptions. However, the model of implementation is not revised subsequent to this chapter, other than to informally point to possible directions for enhancing it in the future. The primary focus is on applying the model of implementation as defined in this chapter, and on evaluating its performance as a score-following component within an operational automated musical accompaniment system.

2.12 Comments on the Validity of the Model

The model for stochastic score following views the score as a sequence of events having a real-valued length and a probabilistic description of the expected sensor output during that event. The lengths of events are interpreted as a measurement along an idealized dimension of distance, relative to other events in close proximity within the score. The location of a performer is characterized by a density function that ideally ranges over the entire length of the score. This estimate can be revised subsequent to receiving new observations from sensors. This revision is defined by a conditional density of the performer's destination position conditioned on three variables—the performer's source position, an estimate of the distance moved by the performer, and the new sensor output. The uncertainty about the current source position of the performer can be incorporated into the update process in a probabilistically sound fashion. This general model can be transformed into a model that is more tractable both for estimating the component density functions and for final implementation on a computer. This simplification is accomplished by assuming that the sensor output is stochastically independent of the source position and estimated distance, that actual reported values may be substituted for observations and distance variables, and that the performer's motion through the score can be adequately approximated by a convolution operation. A discrete model can be derived from the continuous model by sampling all density functions and applying rectangular approximation to calculate the integrals. This discrete model can be implemented in real time by calculating the destination density over only a portion of the score and applying the fast Fourier transform (FFT) to calculate the discrete convolution.

The various models and their derivations introduce several potential sources of error. This may reduce both the accuracy and the precision of the generated estimates of the performer's location, relative to the initial idealized model. To a large extent, several of the assumptions upon which the models are based can be supported by two simple assumptions—that the performer's motion through the score is highly sequential and that successive application of the model occurs frequently thereby insuring a short elapsed time between updates. These assumptions are fairly accurate for the kind of musical accompaniment task considered within this document. The assumptions might be violated if performers tend to miss repeats (initiating a large jump forward in the linear score) or if the performer's timing is strongly influenced by listening to the accompaniment. The extent to which other sources of error influence the accuracy of the tracking system largely depends upon the subsequent definition of the density functions used in the model.

The impact of the component density functions on the model's performance cannot be overestimated. These functions provide the basic ability for any implementation of the model to distinguish the performer's location. If these functions are generally very sharp and readily discriminate between possible destinations of the performer under most circumstances, the model will likely enable accurate tracking even in the presence of nontrivial relative numeric errors. To the extent that under certain conditions and given certain input these densities are relatively smooth and provide little discrimination between score positions, the predictions made by the model will be less accurate and perhaps unacceptable. It is as important to identify the most relevant information upon which to base the density functions as it is to precisely and accurately describe the actual shape of the functions.

It is interesting to consider that in many ways defining the density functions is analogous to the knowledge acquisition process commonly associated with construction of rule-based systems. In both cases, information relevant to the task must be extricated from the less relevant or irrelevant information also available, and must be encoded in a format that allows a software system to apply it. A major difference between these two situations, as is hopefully evidenced in the next several chapters, is that defining density functions not only demands quantification of the information to be applied, but to a certain extent defines *a priori* an appropriate method for so doing. In addition, by having first posited a model of what densities are required and the operations that will be applied to them, it may be possible to theoretically evaluate the adequacy of the defined densities. While this will not eliminate the need for empirical testing of a model, it may help to identify serious shortcomings prior to such testing. It may also significantly enhance one's subsequent understanding of why the testing was or was not successful.

Finally, the density functions provide an explicit representation of the uncertainty that may still exist even after careful estimation of the functions has been completed. A software system may be able to appropriately adjust the decisions it makes or the actions it takes if it can recognize uncertainty in the results generated by the model. It often may be better for the system to recognize that something is not known rather than to simply act with false certainty based upon a questionable conclusion. This is generally true for the designer of the system as well.

Chapter 3

A Stochastic Model of Motion under Positive-valued Rate

3.1 Motivation

A model for stochastic score following was provided in Chapter 2. This model incorporates an estimate of the distance that a performer moves (the amount of score performed) between observations. To use this distance information, the model requires a conditional density function defining the likelihood that the performer has actually moved a certain distance given an estimated distance. Naturally, this density function will depend upon both the inaccuracies of the method used to generate a distance estimate and the nature of a singer's movement through the score. Defining this distance density function for the score-following model requires an empirical investigation of vocal performance.

This chapter presents a stochastic model of motion along a continuum and applies it to vocal performances. First, a theoretical description of the model is provided. This proposed model is partially based upon expected properties of vocal performances. Next, the model is fit to data extracted from actual vocal performances. These performances were given by a number of individuals (primarily vocal students) of various voice parts singing a variety of Western classical pieces. These recordings are thought to comprise an informative sample from the target population. Subsequent to fitting the model to the data, some important properties of the fitted model are empirically demonstrated. Finally, this chapter provides specifics about implementing this model as part of the score-following system. It also re-examines some of the assumptions of the full score-following model within the context of the presented model of motion.

3.2 Properties of Motion under Positive-valued Rate

A singer's motion through a musical score exhibits several important properties. These characteristics of vocal performances must be well characterized by any stochastic model used to describe this motion. The first important property is that the motion is almost always continuous. It is rarely the case that the performer jumps around in the score or stops. This property is especially true of performances by skilled performers like those participating in this study. The data collected for this project included over 60 vocal performances. All of the singers were familiar with the pieces they performed. In this set of performances, there were no examples of jumps or omitted notes, and only 1 example of something that might be viewed as stopping—namely, breathing in the middle of a note. A second important property is that a singer's motion is almost always in the forward direction. A singer certainly never sings in reverse through the score, and backward jumps are very rare. Again, no such instances were observed in the collected performances. Finally, there will be intentional as well as random adjustments to the rate of motion (or *tempo*) during the performance. Such adjustments will appear even in sections of the score where no tempo changes or variations are explicitly indicated.

The actual score distance moved by a performer over a given period of time, $D_{\Delta T}$, will equal the product of the elapsed time, ΔT, and the average rate over that time, $R_{\Delta T}$:

$$D_{\Delta T} \quad = \quad R_{\Delta T} \quad \times \quad \Delta T$$

The unit of measurement for ΔT can be thought of as milliseconds. Either beats or idealized milliseconds (*i.e.*, 1 beat *always* equals x milliseconds) is appropriate for the unit of distance. The unit for rate is defined by the ratio expressing unit of distance over unit of time (*e.g.*, beats per millisecond).

Although an average rate can be calculated for a given period of time, it is highly unlikely that the performer will consistently maintain that rate over any lengthy period. Changes to the rate will manifest as variations between the expected and actual onset times of individual notes. We can represent the rate variations over a period of time as several different rates during subintervals of time:

$$R_{\Delta T} \times \Delta T \;=\; \sum_{i=1}^{n} R_{\Delta t_i} \times \Delta t_i \quad \text{where} \quad \sum_{i=1}^{n} \Delta t_i = \Delta T$$

If for the moment we consider only the case where the performer attempts to maintain a fairly constant target tempo, R_C, then it might be reasonable to expect that the rates over the subintervals, Δt_i, will vary around the target. One way to represent this variation around a target rate is as follows:

$$R_{\Delta T} \times \Delta T \quad = \quad \sum_{i=1}^{n} (R_c \times \varepsilon_i \times \delta_i) \quad \times \quad \Delta t_i$$

Here the ε_i define the effects of random changes to the rate while the δ_i define the effects of the desired, controlled rate changes initiated by the performer. The former type of change may arise due to unintentionally late entrances, unintended slowing of the tempo, or simply motor noise (control noise inherent in the human nervous system and muscles). The latter type of change may result from expressive timing or expressive tempo variations like *rubato*. The formula indicating variations in rate can also be written as follows:

$$R_{\Delta T} \quad = \quad R_c \quad \times \quad \frac{\sum_{i=1}^{n} \varepsilon_i \times \delta_i \times \Delta t_i}{\Delta T}$$

to express the average rate as a scaling of the target rate. Finally, dividing both sides of this equation by the target rate gives:

$$\frac{R_{\Delta T}}{R_c} \quad = \quad \frac{\sum_{i=1}^{n} \varepsilon_i \times \delta_i \times \Delta t_i}{\Delta T} \qquad [3.1]$$

This equation defines the relative error between the average rate over ΔT and the target rate.

Both the random and controlled changes of rate exhibit several important properties. First, since the motion of the performers is continuous and in the forward direction, all ε_i and δ_i are strictly positive. They represent deviations from the target rate as a proportion of that rate. Second, the large majority of the ε_i will be close to one. It is unlikely that a competent performer will randomly shorten or lengthen a note by an extreme amount. Third, in the event that the ε_i alter the rate significantly enough that the performer notices, he or she may attempt a controlled change to compensate. For example, if the performer notices a slowing of the tempo or a late entry, the next note may be appropriately shortened in order to pull the tempo toward the target or to make the entrance at the "expected" time. Finally, the controlled changes made by the performer for expressive purposes are also likely to counteract one another over time. Expressive timing and rubato are used to stretch the underlying tempo but not to completely change it. Note that as previously stated, we momentarily postpone consideration of intentional long-term changes in tempo.

These properties of the error terms have an important effect on equation 3.1. As the number of time intervals, n, increases, the errors weighted by elapsed time will tend to counteract one another in the summation. Over time, the variance of the summation will decrease, and the summation will therefore approach the total elapsed time, ΔT. Since this value appears in the denominator of the error term, over time the error will approach a value of one. This convergence in turn implies that as ΔT increases, the variance of the average rate should decrease, and the average rate should approach the target rate. This decrease in variance agrees with our knowledge about the behavior of a sample average as the size of the sample increases. Namely, the sample average should approach the population average, which due to the controlled nature of vocal performance can be assumed to be very close to the target rate, R_C.

While it is useful to know that the variance of the average rate should decrease over larger periods of time, more specific information about the distribution of the average rate is needed. It can be expected that for a fixed length of time, the error term will be skewed about a value of one—the value at which the average rate equals the target. There are several reasons to believe that a skewed distribution will exist. First, since a performer's rate of motion is always positive, the rate can never assume a value at or below zero. On the other hand, it is possible (though unlikely) that the average rate may assume a value twice that of the target rate. This situation is more likely over short periods of time, like those that will be of most interest for the score-following model.

While the exact nature of the skewness will depend upon the distributions and magnitudes of the underlying sources of error, one possible way to reduce this skewness would be to fit a distribution to the logarithm of the average rate. This transformation also has the advantage that as the value of the average rate approaches zero, the logarithm of the average rate will approach negative infinity. Thus a symmetric, non-truncated density like a normal curve might reasonably approximate the true density of the transformed error. Approximating the logarithm of the average rate by a normal distribution corresponds to approximating the average rate by a lognormal distribution. This distribution is defined as follows:

$$\frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

where μ and σ represent the mean and variance of the underlying normal distribution of the logarithm of x. Note that if the median of the lognormal distribution is 1, then the median of the underlying normal distribution will be 0. Since the mean and median of a normal distribution are identical, the mean of the underlying normal distribution in this case will be $\mu = 0$. Also, the mean of a lognormal distribution is always greater than the median, due to the positive skewness of the density. Given a normal distribution with mean μ and variance σ^2, the mean of the corresponding lognormal distribution is $e^{\mu+\frac{1}{2}\sigma^2}$.

A more compelling reason for use of a lognormal distribution comes from the behavior in the limit of random proportionate changes. If we express the average rate over a fixed time as the result of many random, and fairly equal, proportionate changes to the target rate:

$$R_{\Delta T} \quad = \quad R_c \times \varepsilon_1 \times \varepsilon_2 \times \varepsilon_3 \times \cdots \times \varepsilon_n$$

then the difference between the logarithms of the average rate and the target rate will be a sum of several random variables:

$$\ln R_{\Delta T} - \ln R_c \quad = \quad \ln \varepsilon_1 + \ln \varepsilon_2 + \ln \varepsilon_3 + \cdots + \ln \varepsilon_n$$

As the number of random proportionate changes increases, the sum of their logarithms will likely approach a normal distribution. For a relatively long period of time, it is very likely that this distribution will be well approximated by a normal distribution. Depending upon the distributions of the random proportionate changes, a normal approximation may be reasonable even for relatively short time periods. This property of random proportionate changes is commonly referred to as *the law of proportionate effect*, and is often applied to model positive-valued data in fields such as biology, economics, and geology. A good historical overview of the lognormal distribution and its application is provided by Aitchison and Brown (1957).

Given a normal approximation to the distribution of the logarithm of the average rate over a fixed time period, and treating elapsed time as an accurately measurable constant, an estimate of the distance moved can be obtained. Consider that:

$$\ln\left(\frac{D_{\Delta T}}{R_C \times \Delta T}\right) \quad = \quad \ln\left(\frac{R_{\Delta T} \times \Delta T}{R_C \times \Delta T}\right) \quad = \quad \ln\left(\frac{R_{\Delta T}}{R_C}\right)$$

If ln $R_{\Delta T}$ is assumed normally distributed with mean $\mu = \ln R_c - \frac{1}{2}\sigma^2$, then correspondingly ln $D_{\Delta T}$ is also normally distributed with a mean of $\mu = \ln R_c - \frac{1}{2}\sigma^2 + \ln \Delta T$. This assumes that the mean of the lognormal distribution describing $R_{\Delta T}$ is R_C—that is to say, the target rate is the expected rate. Note that while the mean of the logarithm of the average rate remains unaltered over time, the mean of the logarithm of the distance increases in proportion to the time. However the variance of both logarithms behaves identically and is not affected by the term ln ΔT. The lognormal density approximating the conditional density of actual distance conditioned on an estimated distance of $d = R_c \times \Delta T$ will be defined by the following equation:

$$f_{I-J\,|\,\mathrm{D}}\,(i-j|\; R_C \times \Delta T) \quad = \quad \frac{1}{(i-j)\sigma\sqrt{2\pi}}\, e^{\frac{-\left(\ln(i-j) - \ln R_C + \frac{1}{2}\sigma^2 - \ln \Delta T\right)^2}{2\sigma^2}} \qquad \textbf{[3.2]}$$

This is the function that is fit to the vocal performance data for various ranges of elapsed times, ΔT.

One problem regarding the use of the suggested lognormal distribution is that the performer's target rate, R_C, will probably not be known in advance. However, if the stochastic tracking system is able to adequately track the performer's score position over time, the target rate can be estimated by an average rate calculated over several previous position estimates. Since the average rate is expected to converge to the target rate over longer periods of time, a past average over a sufficiently long period of time is likely to be a good estimate of the target. In the case of a good estimate of the target, the error properties previously discussed would be equally valid for a ratio of an average rate to the estimate. Furthermore, these error properties could directly apply for the ratio of average rate in the immediate future to average rate in the recent past. Thus, the logarithm of this relative error also will be approximately normally distributed. It is therefore appropriate to estimate this latter density for use in the score-following model. As will be shown, this can be done by examining pairs of successive average rates appearing within actual performances.

The question remains as to how gradual or sudden changes in the underlying target rate will impact estimates of the distance density function. In general, this will cause our estimate of R_C—the average rate over an immediately preceding region of score—to be less accurate. As a result, we will be approximating a distribution with one mean by a distribution with a slightly different mean. If for short periods of time, like those of most interest for score following, these changes are gradual enough so as not to be much more significant than expressive timing or *rubato*, then our distribution may not be affected. If such changes are large and sudden, however, we should expect to observe them as outliers in the data. These deviations might alter our approximated normal density.

To the extent that these sudden changes are explicitly indicated in the score, we might be able to define a different distance density function for use when we believe the performer to be in these regions of the score. Alternatively, if we can approximate the likely changes in tempo by proportionately modifying the length of successive events in those regions, the distance density may not be significantly altered from the distance density in regions where a more constant target tempo is expected. If one wants a locomotive to spend more time moving between two train stops, equally effective solutions are to slow its rate or to lay down more track. For sudden or large tempo changes that are not explicit in the score, the score-following model will simply have to rely on the observation densities to clearly distinguish position. The observations likely would be critical in these situations even when using truly accurate distance densities, since large tempo changes that are not marked in the score will occur rarely and could not be anticipated without significant additional modeling of vocal performance. Observations will also be critical for determining score position in other rare situations that either cannot be anticipated or are not explicitly modeled. These situations include cases when performers repeat a note already sung or skip measures because they miss repeats, misread the music, or wrongly sequence the pages of the score.

The following summarizes the stochastic modeling of distance traversed by a performer and the expectations about the distribution:

1. A singer's motion through the score is almost exclusively continuous and in the forward direction (*i.e*., has positive-valued rate), and is subject to unintentional as well as intentional changes in rate.

2. In regions where an underlying target rate may be assumed, the average rate of the performer will be roughly lognormally distributed with a mean equal to the target rate, R_C. The distribution of the logarithm of the average rate will be roughly normal with a mean equal to $\ln R_c - \frac{1}{2}\sigma^2$. The distribution of average rate will have a variance that is inversely proportionate to the length of time over which the average rate is calculated.

3. The current target rate, R_C, can be adequately approximated by an average rate calculated from multiple previous position estimates spanning a sufficiently long period of time. If the approximation is good, then the ratio of an actual average rate to a predicted average rate will also be approximately lognormally distributed.

4. Treating elapsed time as an accurately measurable constant, the distance density required for the stochastic score-following model can be obtained directly from the density describing rate. The distance density also will be roughly lognormal with a mean equal to the product of the target rate and the elapsed time, $R_c \times \Delta T$. The distribution of the logarithm of the distance will be roughly normal with a mean equal to $\ln R_c - \frac{1}{2}\sigma^2 + \ln \Delta T$. The variance of this distribution will be equivalent to the variance of the distribution describing logarithm of rate.

Subsequent examination of vocal performances will be used to evaluate this model, as well as the validity of the convolution assumption used to simplify the general score-following model.

3.3 Estimating the Distance Density Using Actual Vocal Performances

For purposes of determining the density functions required by the score-following model, 61 live vocal performances were recorded in a laboratory setting. Subsets of these performances were used to obtain the subsequently described data, as well as the data presented in other chapters of this document. A total of 13 performers participated in the study. All performers had at least some university level vocal training in Western classical music. The experience of the performers ranged from sophomore university students minoring in voice to professional opera performers. The voice parts of the performers consisted of 3 sopranos, 2 mezzos, 1 contralto, 1 countertenor, 2 tenors, 3 baritones, and 1 bass for a total of 6 female performers and 7 male performers. This is believed to be a reasonable constituency of voice parts for the population of performers and the type of music targeted by this study.

The performers were recorded using directional microphones placed in close proximity, so as to capture the vocal performances in isolation rather than with accompaniment. All performances used a live accompanist. In many cases, the accompanist and vocalist had not performed together prior to participating in this study, but were permitted one rehearsal of each piece prior to recording. The vocalists were familiar with all pieces they performed, which frequently included works currently studied as part of their semester training. The pieces were made available to the accompanist for private rehearsal well in advance of the recording sessions. Thus for the most part, the data presented in this document can be taken to represent the behavior of vocalists when both they and their accompanist are relying mainly on general musical background and experiences along with individual rehearsal of what is explicitly notated in the selected score. The nature of such performances may or may not be in contrast to highly rehearsed performances where more specific experiences might be applied.

Of the 61 performances recorded, 15 were made of the well-known tune "Happy Birthday", purely for the sake of having some directly comparable data across a number of performers should this later prove desirable. The remaining 46 recordings included 23 different works. Each performer was asked to perform 2 to 3 different pieces of their own choosing and, time permitting, two recordings were made of each piece. While the mapping between recordings and pieces is not as exact as the 46-23 ratio would suggest, many performers were recorded twice for every piece they performed.

The 23 pieces included 8 songs from the romantic and contemporary periods, 3 classical arrangements of folk songs and spirituals, and 12 arias from operas, oratorios, and cantatas spanning the baroque, classical, romantic, and contemporary periods. The pieces included 10 in English, 8 in Italian, 5 in German, and 1 in French (one performer sang an English translation of a German piece sung by other performers, hence these numbers total 24). This distribution of languages (and possibly the distribution of musical styles as well) is biased by the fact that performers were requested to select pieces in a variety of languages with at least one piece in English. A piece in English was explicitly requested since this study included plans to use observations based on phonetic content of the scores. Otherwise, Italian and German would likely have dominated the language content of the data. Requesting pieces in English was a simple way to increase the likelihood of more balanced and comprehensive phonetic content.

The data presented in this section represents a subset of 20 recordings. The other 26 recordings were set aside both to prevent use of multiple recordings of the same performer singing the same piece (which might bias the estimated statistics) and to preserve some unexamined recordings for testing of the system. The 20 recordings contained 2 performances by each of 10 performers and encompassed 16 different songs. Any multiple performances of the same song were given by different performers. Of the 10 performers, 4 were female and 6 were male with all available voice types represented. Note that this group does not include every performer who participated in the recording sessions. With few adjustments, this subset of performances is the source of all data used to estimate the density functions required by the score-following model. Throughout this and subsequent chapters, it will often be referred to as either *the sample of 20 performances* or simply as *the sample of 20*.

The distance density to be modeled is based on fitting a lognormal distribution to the actual average rate divided by the predicted average rate, where the predicted rate is an average rate observed in the recent past. In order to estimate this density from the performances, the following approach was used. First, it was necessary to obtain timing information for each performance by identifying the time at which each note in the score was initiated by the performer. This time is referred to as the *onset time* of the note. The onset time was defined to be the point where significant pitched sound corresponding to a note first appeared in the sound signal. Since vocal students are often encouraged to produce the pitched vowel of a syllable simultaneous with the beat (thereby producing any leading unvoiced consonants prior to the beat), selecting the start of the pitched sound to indicate the note onset seemed appropriate.

Pitched portions of the performance were identified by pitch detection software that accepts digitized sound signal and produces a time-stamped pitch estimate for roughly every 100 ms portion of the signal that exceeds a constant amplitude threshold. This segmentation of the signal resets whenever at least 15 ms of signal remain below the threshold. This duration is just longer than the duration of a single pitch period for the lowest note in the bass range. No pitches are reported in regions where the signal remains entirely below the threshold, or when a block of signal above the threshold is less than 90 ms. A sequence of time-stamped pitch readings was generated for each performance, and the sequences were parsed by hand to determine the start of each note. Parsing was based on changes in pitch and the appearance of unpitched gaps in the output. In some cases, accurate parsing required use of a tool that displays a graph of the sound waveform and permits playback of selected regions of the signal.

Having obtained note onset times for all 20 performances, pairs of *predicted average rate* (used to estimate the target rate, R_C) and *actual average rate* were calculated. An average rate is calculated by dividing the distance between two notes, as indicated in the score, and the actual elapsed time between those same notes in the performance. Thus, a rate of 1 means the performer sang at the nominal tempo in the score, a rate above 1 means the performer sang at a faster tempo (*i.e.*, covered that portion of score in less time), and a rate below 1 means the performer sang at a slower tempo (*i.e.*, covered that portion of score in more time). A predicted average rate was calculated over regions of score starting at each note in the score up to the first subsequent note with an onset time at least 2 seconds later. Thus all predicted rates were calculated over a minimum time period of 2 seconds. Since most notes have a duration less than 2 seconds and many have a duration less than 1 second, a period of 2 or more seconds generally will include multiple notes, reducing the variance of the predicated rate relative to the true target rate. An actual average rate was calculated over time periods marked by each subsequent note whose onset time fell within 3 seconds of the region used to calculate the predicted rate. Thus for each predicted average rate, several pairs of predicted and actual rates were generated where the elapsed time for each actual rate calculation spanned less than 3 seconds. A graphic example of these rate calculations is presented in Figure 3-1.

Pairs of rates were not calculated across sudden tempo changes that were explicitly indicated in the score. Rate pairs were calculated up to the tempo change and then were calculated starting from the tempo change as if it were the beginning of a new piece. The only exceptions were notes or rests marked by a *fermata*—a mark indicating that the performer is allowed to sustain the note or rest indefinitely. These notes were not used for any rate calculations. Rate pairs were calculated up to the start of the fermata and calculations resumed with the note immediately following the fermata. Sections of score that contained marks to indicate gradual tempo changes were not excluded from the rate calculations. 8 of the 16 songs appearing in the sample of 20 contained at least 1 sudden tempo change or fermata, and no single piece contained more than 4. Finally, rate pairs were discarded if the elapsed time for the predicted rate spanned more than 5 seconds, since larger time spans generally encompassed a long rest between two notes in the vocalist's part. The accompanist, not the vocalist, controls the tempo during these periods.

Recall that if the rate ratio for a fixed elapsed time is lognormally distributed, the logarithm of the ratio will be normally distributed. The maximum likelihood estimate of the parameters μ and σ^2 for a lognormal distribution are given by the mean and variance, respectively, of the logarithms of the data points. This approach was used to fit lognormal distributions to the rate pair data. Statistics can be applied to assess the lognormality of the data sets. The mean and variance of the lognormal distributions can be calculated from the parameters μ and σ^2. These values were calculated and compared against the predictions that the mean will be 1 and the variance will exhibit an inverse linear relationship to the

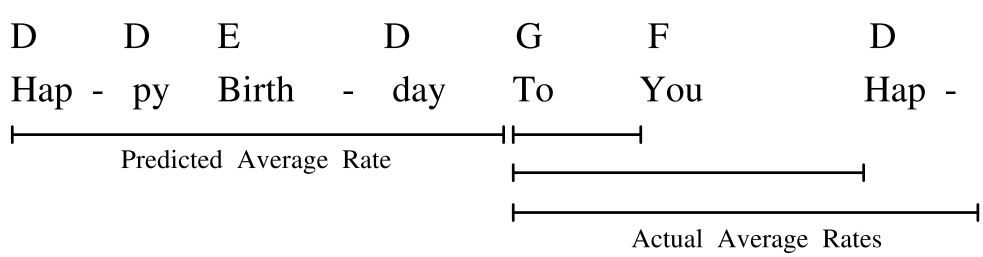


Figure 3-1. For each predicted rate that is calculated, multiple actual rates are calculated over regions of score associated with successive note onset times.

elapsed time. Lognormal distributions defined by a mean of 1 and variance obtained from a fitted regression line can be compared against the data sets, again to assess goodness of fit. The results of calculating and applying these estimates and assessments are detailed in the remainder of this section.

Once all rate pairs were calculated, the difference between the logarithms of the two rates in each rate pair was calculated:

$$\ln\left(\frac{R_{Actual}}{R_{\Pr edicted}}\right) = \ln R_{Actual} - \ln R_{\Pr edicted}$$

These values were grouped by ranges of the elapsed time over which the actual rate was calculated. Partitioning according to actual elapsed time was necessary since it is expected that the variance will change relative to this value. Means, standard deviations, and standard errors for the means (an approximation to the standard deviation of the distribution of the mean) were calculated for data sets with elapsed time ranges of 200 ms centered around 600, 900, 1200, 1500, 1800, 2100, and 2400 ms. Prior to calculating these values, the histograms of the data sets were examined for extreme outliers. These data points were identified and discarded, since even a few such outliers can drastically impact calculation of the standard deviation. The outliers resulted from performers sustaining or shortening notes in extreme ways, in some cases essentially adding a fermata or altering the rhythm indicated in the score. No more than one-half of one percent of any individual data set was discarded. The values calculated from the data sets are presented in Table 3-1, along with the number of discarded outliers.

Table 3-1. Statistics for samples of ln $R_{\Delta T}$ – ln R_C having actual elapsed time within 100 ms of the stated elapsed time.

Elapsed Time	Sample Size	Mean	Standard Deviation	Standard Error	$\mu + 0.5\sigma^2$ / S.E.	Discarded Outliers
600 ms	762	-0.005	0.227	0.008	2.6	2
900 ms	546	-0.013	0.192	0.008	0.7	2
1200 ms	507	-0.023	0.159	0.007	-1.5	2
1500 ms	508	0.007	0.149	0.007	2.6	2
1800 ms	458	-0.006	0.130	0.006	0.4	2
2100 ms	520	0.008	0.131	0.006	2.8	1
2400 ms	423	-0.012	0.119	0.006	-0.8	2

The smallest range of values for any of the data sets is on the order of ±0.5. In comparison to this value, the means for all data sets are relatively close to 0. The average of the means for the seven data sets is -0.006, and the most extreme deviations from zero assume negative values. In addition, only two of the data sets have a positive mean. Because lognormal distributions are positively skewed, the mean of the distribution always exceeds the median. The median of the underlying normal distribution is the logarithm of the median of the lognormal distribution. This value is also the mean of the underlying normal distribution. It is less than the logarithm of the mean of the lognormal distribution. Consequently, since the examined data sets contain the actual rate minus the predicted rate, negative means are to be expected if the actual rate is lognormally distributed and the predicted rate is a reasonable approximation to the expected actual rate. Also as expected, the standard deviations calculated from the data sets are generally decreasing as elapsed time increases, except between 1800 and 2100 ms, indicating convergence to the mean. In other words, the actual average rate appears to converge to the predicted average rate as elapsed time increases.

The mean of a lognormal distribution is related to the parameters of the underlying normal distribution by $e^{u+\frac{1}{2}\sigma^2}$. Assuming that the actual rate is lognormally distributed with a mean equal to the predicted rate, the means of the logarithmic data sets should equal negative one-half the variance. The sixth column in Table 3-1 displays the mean plus one-half the variance divided by the standard error. As shown, 3 out of 7 of the data sets have a mean that is beyond 2 s.e. from negative one-half the variance. Under an assumption that the distribution of the estimated mean is normal, a 95% confidence interval is given by 2 s.e. on either side of the estimated mean. If a random set of values was selected from a truly lognormal distribution and the logarithms calculated, the transformed data would produce an estimated mean that exceeds negative one-half the true variance by 2 s.e. only 5 percent of the time. While such a comparison is not strictly reasonable in this case, since the variance was also estimated from the data, the differences between the two estimates provide one measurement for assessing how well the fitted lognormal distributions approximate the data sets. Figure 3-2 provides a graph of the means of the lognormal distributions fit to the rate pair data sets. These values are close to 1 with an average of 1.007.

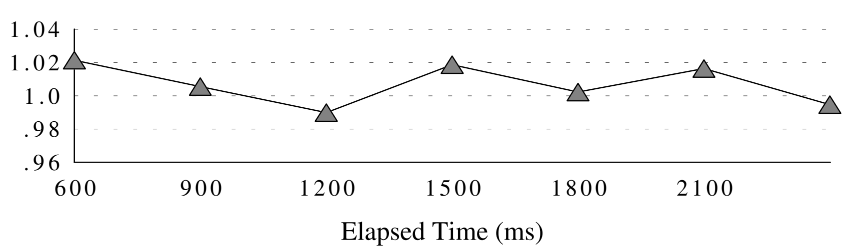


Figure 3-2. Graph of the means of the lognormal models fit to the rate pair data.

The variance of a lognormal distribution can be calculated from the mean and variance of the underlying normal distribution as follows:

$$\beta^2 \quad = \quad e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right) \qquad [3.3]$$

Using the standard deviations and means calculated from the data sets, the variances were calculated for the corresponding lognormal distributions of actual rate divided by predicted rate. Just as the variance of a sample mean is inversely proportionate to the sample size, these variances are expected to be inversely proportionate to the elapsed time over which the actual average rate is calculated.

Figure 3-3 contains two graphs that examine a simple regression line fit to the inverse of the calculated lognormal variances. The upper graph displays the simple regression line and the inverse variances. The lower graph contains the residual errors for each of the points against the line. The regression line appears to be a good fit. The residuals are fairly random and are largest where the sample standard deviations are not decreasing, between 1800 and 2100 ms. The slope of the fitted line is .02796 and the intercept is 2.64.

Given the regression line for the inverse variance of the lognormal distributions over elapsed time, and assuming that the mean of the lognormal distribution equals 1, a lognormal distribution can be estimated for any elapsed time. This estimation is accomplished by first calculating the variance of the lognormal distribution from the regression line, and then substituting that value along with $e^{\mu+\frac{1}{2}\sigma^2}=1$ into equation 3-3. This yields an expression for σ that can be evaluated and subsequently used to calculate μ.

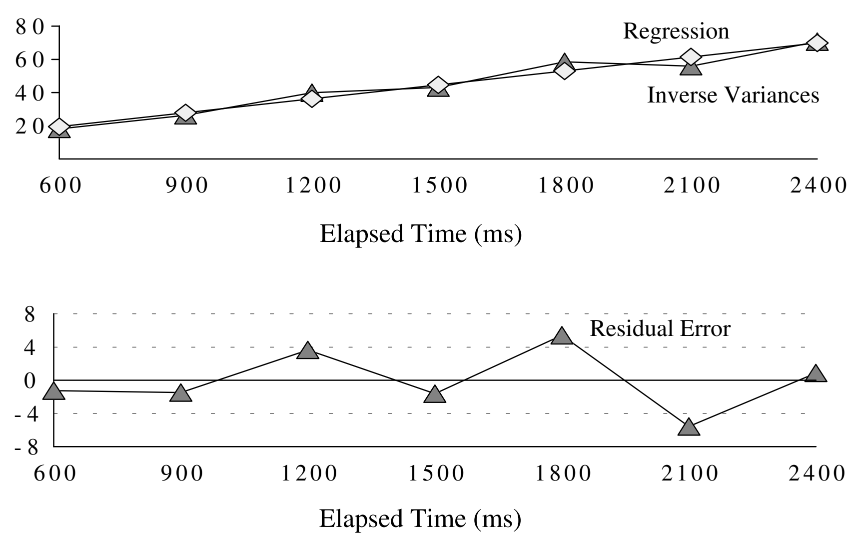


Figure 3-3. Two graphs examining a simple regression line against the inverse of the variances for lognormal models fit to the rate pair data. The upper graph shows a regression line fit to all data points; the lower graph shows the residual errors of the points against the line.

The Kolmogorov-Smirnov statistic (or K-S statistic) is a method for assessing how well a proposed cumulative distribution function models a set of data. It is based on the largest difference between the cumulative distribution of the data and the value of the proposed cdf at the corresponding point. For a random sample of given size, an approximation can be derived for the likelihood that the sample distribution will vary from the true underlying population distribution. This value can be used to provide confidence levels for rejecting a proposed distribution function. Table 3-2 provides values of the K-S statistic when the samples of differences of the rate logarithms were fit to normal distributions. The means and variances were calculated from the regression line as previously described. For the set of samples used in the regression, the values of the K-S statistic are fairly large compared to what one might expect if the distributions were truly lognormal with mean and variance as calculated.

The large values of the K-S statistic might indicate that the distribution of a small number of intentional and accidental rate changes on the part of the performer is not exactly lognormal. Samples for longer elapsed time periods were subsequently calculated and the values of the K-S statistic for these samples are also displayed in the table. Again, this test considered a normal curve of mean and variance derived from the previously calculated regression line. Approximately 1 percent of each of these samples

Table 3-2. K-S statistics for samples of $\ln R_{\Delta T} - \ln R_C$. The samples were fit to a normal distribution with mean and variance derived from the regression line fit to the inverse calculated lognormal variances. The P-value indicates the probability of observing a K-S statistic at least as large as the one actually observed, if the true distribution were normal.

Elapsed Time	Sample Size	K-S Statistic	P-Value
600 ms	762	0.0854	< 0.01
900 ms	546	0.0596	< 0.05
1200 ms	507	0.0659	< 0.05
1500 ms	508	0.1121	< 0.01
1800 ms	458	0.0410	> 0.10
2100 ms	520	0.0759	< 0.01
2400 ms	423	0.0505	> 0.10
3000 ms	417	0.0389	>> 0.10
4000 ms	388	0.0475	>> 0.10
5000 ms	412	0.0520	> 0.10

were discarded as noticeable outliers prior to this analysis. The number of recognizable outliers appears to increase slightly over time. There is noticeable improvement in the K-S statistic for these samples. Such gradual improvement for larger elapsed times might be explained by the law of proportionate effect as previously discussed.

The distributions for the samples in this study might also be affected by high-level patterns in the 20 performances. Since the data points in the samples include nearly all notes in every performance, any recurring phrases or patterns in the music that are performed with consistent tempo deviations might bias the samples or invalidate an assumption of randomness. Also, rates that are calculated from overlapping regions in the score will be similarly influenced by tempo changes or shortened notes that occur in the shared section of score. A model that explicitly accounts for rate changes might produce better estimation results.

Finally, several aspects of the method used to extract the data sets from the performance may have affected their distributions. First, the data sets contained rates calculated only at note onsets. In addition to the difficulty of accurately identifying the onsets, using only rates calculated at onsets may augment the effects of high-level patterns in the scores. For instance, if repeated rhythmic patterns in a score, in combination with a performer's chosen tempo, yield many notes with an idealized duration of 700 ms; then only the instances where a note is shortened would contribute to the data set centered around 600 ms. Using a window as large as 200 ms centered around the target elapsed time may also adversely affect the variance estimates. Similarly, using a smaller time window for calculating the predicted rate might improve the estimates. One way to eliminate these problems might be to collect a larger set of performances and use only a limited portion of each performance. If rates were calculated at points randomly selected from within each performance, equally large data sets with less bias might be obtained.

However, the rate data contained in the examined samples provides some important information relative to modeling the motion of vocal performers. First, the variances of the lognormal distributions fit to the ratios of actual to predicted average rate appear to be inversely proportionate to the elapsed time over which the actual average rate is calculated, at least for the range of times examined. This inverse linear relationship was expected since the variance of an average decreases in proportion to the number of samples. Observing this behavior supports the belief that the actual average rate varies less from the predicted average rate as elapsed time increases. Second, the samples examined do not refute the assumption of a mean less than 0 for the logarithm of the rate error, which could correspond to a mean of 1 for a lognormally distributed ratio of actual to predicted average rate. The sample means do exhibit a slight positive bias relative to the predicted means. However, average rate in the recent past appears to be a good estimate of the average rate in the immediate future. Finally, while lognormal densities with mean and variance derived from the calculated regression line are not bad approximations, tests of fit comparing these densities to the samples indicate possible room for improvement. This situation may have occurred because the lognormal distribution is only an asymptotic approximation, because biases are introduced by high-level patterns in the performances, or because of the methods used to analyze and partition the data. Better models of the underlying sources of individual tempo deviations might also suggest ways of improving the overall rate model, as might models of the correlation between the errors of rate estimates derived from overlapping regions of score.

Nevertheless, the fitted lognormal model provides a set of formulas for estimating the distribution of the actual distance moved by the performer given an elapsed time and a predicted rate. First, the variance of the lognormal distribution is approximated via the fitted regression line:

$$\frac{1}{\beta^2} = 0.02796 \times \Delta T + 2.64 \qquad [3.4]$$

Next, if the mean of the lognormal distribution is 1, $e^{\mu + \frac{1}{2}\sigma^2} = 1$ can be substituted into equation 3.3 relating variance of the lognormal distribution to the variance of the underlying normal distribution, and the following formula can be used to determine the variance of the normal distribution:

$$\sigma^2 = \ln(\beta^2 + 1) \qquad [3.5]$$

Finally, this estimate of the variance, along with a mean defined by $\mu = \ln R_c - \frac{1}{2}\sigma^2 + \ln \Delta T$ as previously mentioned, will complete the definition of the distance density specified in equation 3.2. This density can be generated for various values of R_C and ΔT, and the resulting function can be sampled during application of the stochastic score-following model.

3.4 A Model Providing Consistent Estimation When Elapsed Time Varies

The distance density proposed for use in the stochastic score-following model has two important characteristics. First, the model accommodates tempo changes in the performance. It adjusts the likelihood of actual distance moved by a performer according to a short-term prediction of the performer's current average tempo. This performance-specific adjustment to the distance estimate is likely to be more useful in discriminating position than would a model that assumed an unvarying target tempo or that indiscriminately averaged across all possible tempi. Second, the model can provide updated score position estimates after arbitrary time intervals. The density over actual distance varies according to the time that has elapsed since the last update of the score position density function. This time dependency can be used to relax the requirements placed on the frequency of model updates, and subsequently on the frequency with which observations are made. While the first characteristic is relatively straightforward, this second property is worthy of further investigation.

First, it is important to point out that elapsed time, as indicated by ΔT in the distance density, always corresponds to elapsed time relative to the performance, or more generally the input signal used to generate observations. This time may or may not correspond identically to the computation time that has elapsed between updates of the model. Since the musical accompaniment application has strict real-time requirements, for this case it will be fine to associate elapsed computation time with the variable ΔT. However, when using the model under circumstances where real-time constraints may be more relaxed or not a concern, elapsed time as represented by the input signal may not be identical to elapsed computation time. Examples of score following that do not require real-time tracking include analyzing recorded performances, studio post-processing such as correcting pitch and aligning recordings of multiple performers recorded individually, and investigating error propagation in the score-following model by sampling the density function at a very high rate.

Now ideally, the distance density should allow the model to be used under a variety of position estimation scenarios. The frequency with which the position density is updated may depend upon the type of observations made or the methods by which the observations are generated. There should exist some consistent relation between the behavior of the model when updated every 100 ms versus the behavior when updated every 200 ms, etc. It is perfectly valid to ask whether updating the model once after 200 ms is identical to updating the model after each of 2 successive 100 ms periods. Likewise, the model should exhibit some consistency when successive updates occur at different elapsed time intervals. It is equally valid to question the result of first updating the model after 100 ms has elapsed, then subsequently after 300 ms has elapsed.

There are a number of practical applications for the answers to such questions. Certainly, it is important to know whether or not performance of the accompaniment system is altered by the selection of model update frequency. It may also be useful to alter the update frequency according to the computational speed of various computing platforms. Additionally, one might wish to vary the time interval between model updates to deal with observations that either are not available at regular intervals or that cannot be made reliably at a fixed frequency. For example, fundamental pitch is a prime candidate for use in musical score following. However, pitch information is not consistently available during a vocal performance. Most notably, there are periods of silence and periods where the performer produces consonants that are not pitched. Rather than tackling the onerous task of modeling the occurrence and characteristics of all portions of the performance, it would be simpler to model just pitch and then update the model only when pitch information is available or can be reliably extracted from the performance.

In this last example, however, potentially useful information contained in the signal is being discarded because of a simplified model. Recognizing that no pitch value is present in the signal may be helpful in distinguishing score position. Clearly then, altering the frequency of model updates can be partitioned into two scenarios. In the first scenario, observations are always generated so that no portion of any input signal is ignored or discarded, but the amount of signal processed to produce any given observation will vary. Since the behavior of the model as elapsed time varies will then depend upon the definition of the observation density, characterizing this behavior becomes quite complex. Furthermore, altering the amount of signal that contributes to an observation technically may violate an assumption of independence between observations and distance. As discussed in the previous chapter, this independence assumption is used to support the final score-following model. Since the score-following model described simply does not support such timing variations and timing variations can be avoided in practice, there will be no further investigation of varying the time period over which an observation is generated.

The second scenario for altering elapsed times allows for certain portions of input not to be associated with a generated observation. These portions of input therefore should not affect the score position density function generated by the model. Ideally, omitting an observation and then updating the model when the next observation is available should be equivalent to first updating the model with a non-informative, uniform observation density followed by an update incorporating the new observation. More specifically, based on equation 2.16, the convolution integral calculated in the latter instance should be equivalent to the two successive convolutions calculated in the former instance:

$$\int_{j=0}^{\|Score\|} f_{H-J\,|\,D}(h-j|\ d) \cdot f_{\mathrm{Pr}ior}(j)\,\partial j \quad = \quad \int_{i=0}^{\|Score\|} f_{H-I\,|\,D}(h-i|\ \tfrac{d}{2}) \left[\int_{j=0}^{\|Score\|} f_{I-J\,|\,D}(i-j|\tfrac{d}{2}) \cdot f_{\mathrm{Pr}ior}(j)\,\partial j\right] \partial i$$

Note that *h* represents the final location of the performer, while *i* represents an intermediate position of the performer between *j* and *h*. Also, since in the first instance the only information contributing to the update of the position density is the estimate of distance moved by the performer, our estimate of rate will not likely be altered prior to computation of the second convolution. The initial rate estimate will be assumed self-confirming under convolution, although this may ultimately depend upon how the accompaniment system determines the performer's position. Estimates of distance for the first instance change according to elapsed time, which in this case is cut exactly in half. However, any arbitrary division of the time should be permissible, and all possibilities of multiple, successive convolutions should yield equivalent position densities. As a simplification, the fact that convolution is associative means that verification of:

$$f_{H-J\,|\,D}(h-j|\ d) \quad = \quad \int_{i=0}^{\|Score\|} f_{H-I\,|\,D}(h-i|\ \tfrac{d}{2}) \cdot f_{I-J\,|\,D}(i-j|\tfrac{d}{2})\,\partial i$$

will also verify the previous equivalence.

The question of interest now becomes whether the proposed set of lognormal densities, with underlying normal distributions characterized by mean and variance as calculated from the given regression line, will satisfy these kinds of convolution equations under an assumption of constant rate. For example, does self-convolution of the lognormal distance density for $R_C = 1$ and $\Delta T = 600$ yield the lognormal distance density for $R_C = 1$ and $\Delta T = 1200$? For many density functions, this would be a simple matter of calculating the Fourier integral of the function (which statisticians call the *characteristic function* of the density) and examining how it behaves under multiplication. This is true since the convolution of two functions is equivalent to the inverse Fourier transform of the product of their individual Fourier transforms. Unfortunately for us, however, the Fourier transform of a lognormal density cannot be computed in a closed form. While "convolution" properties using a ratio of two lognormal variates instead of a difference can be derived from additive properties of the underlying normal distributions, no closed form expression for the distribution of a sum of lognormal variates is known. However, there is agreement that in practice a sum of lognormal variates can be well approximated by another lognormal variate (Schwartz and Yeh 1982; Beaulieu, Abu-Dayya, and McLane 1995).

In order to provide some support for believing that the lognormal densities from the fitted regression line exhibit the proper behavior when elapsed time varies, repeated convolutions involving these densities were numerically calculated. The density functions were sampled along the time dimension at intervals of 2 ms. The convolution integral was calculated by directly multiplying the two sampled functions and applying Simpson's rule to approximate the integral. Double-precision floating point arithmetic was used for all calculations.

The lognormal density for $\Delta T = 600$ and $R_c = 1$ was numerically convolved three times in succession—thus the function was used a total of four times as part of three convolution integrals. The mean, median (exponentiation of the mean of the underlying normal curve), and variance of the result of each convolution were numerically calculated. The results of this experiment appear in Table 3-3. Note that the actual means and medians agree with the predicted values, but the actual variances are noticeably smaller than the predicted variances. The actual variances are proportionate to elapsed time, and the variance of the curve with mean at 2400 ms is exactly double that of the curve with mean at 1200 ms. The variances for this example appear to satisfy the property $\frac{\beta_{\Delta T2}}{\beta_{\Delta T1}} = \frac{\Delta T2}{\Delta T1}$. This property also held in several additional simulations using other values for ΔT. The difference between actual and predicted variances becomes much worse as the value of ΔT is reduced. Small elapsed times require the fitted regression line to extrapolate beyond the elapsed times encompassed by the data sets. Such extrapolation will be necessary in order for the score-following model to be applied more frequently than once every 600 ms.

Table 3-3. Results of numerical convolution using the lognormal density for $\Delta T = 600$ and $R_C = 1$.

Conv.	Actual Mean	Pred. Mean	Actual Median	Pred. Median	Actual Var.	Pred. Var.
1	1200.01	1200	1184	1184	37083	39853
2	1800.02	1800	1784	1783	55625	61158
3	2400.02	2400	2384	2383	74166	82375

The question that now arises is whether or not the behavior of the variance for lognormally distributed distance, as observed under numerical convolution, can be reconciled with the observation that the variance of actual average rate divided by predicted average rate exhibits an inverse linear relationship to elapsed time. From equation 3.3 specifying the variance of a lognormal distribution, we note that the variance is the mean squared times $e^{\sigma^2} - 1$. This equation allows us to specify the relationship between the variance of the rate ratio density and the variance of the distance density as follows:

$$\beta_{R_{\Delta T}/R_C} = e^{\sigma^2_{\Delta T}} - 1 = \frac{(R_C \cdot \Delta T)^2}{(R_C \cdot \Delta T)^2} \cdot \left(e^{\sigma^2_{\Delta T}} - 1\right) = \frac{1}{(R_C \cdot \Delta T)^2} \cdot \beta_{D_{\Delta T}}$$

Note that the mean of the rate ratio is 1 and the mean of the rate density is taken to be R_C. Using the given equivalence in combination with the observed behavior of the variance of lognormal densities under convolution:

$$\frac{\beta_{D_{\Delta T2}}}{\beta_{D_{\Delta T1}}} = \frac{\Delta T2}{\Delta T1}$$

we obtain the following requirement for the rate ratio variances:

$$\frac{\beta_{R_{\Delta T2}/R_C}}{\beta_{R_{\Delta T1}/R_C}} = \frac{(R_C \cdot \Delta T1)^2}{(R_C \cdot \Delta T2)^2} \cdot \frac{\Delta T2}{\Delta T1} = \frac{\Delta T1}{\Delta T2}$$

Thus the proportion of the variances of the rate ratios must be the inverse of the proportion of the elapsed time. If the inverses of these individual variances must also be linear in elapsed time, then the intercept of the line relating inverse variance and time *must* be 0. The slope, however, may assume any value.

Now the intercept of the line fitted to the inverse variances is 2.64, but the standard error of this calculated intercept is 4.10. Thus an intercept of zero is well within 1 s.e. of the fitted intercept. By assuming an intercept of 0 and applying linear least squares to obtain an estimate of the slope of a line through the inverse variances from the data sets, a value of 0.02948 is obtained. Our initial estimate of the slope was 0.02796, and the standard error of this estimate is 0.00254. Thus our new estimate is within 1 s.e. of the initial estimate calculated from the data. This revised line should now provide a set of lognormal densities that behave as desired under convolution. Table 3-4 shows the results of numerical convolution of the lognormal density for $\Delta T = 600$ and $R_C = 1$ generated from the revised line. These results are much improved from those of the initial simulation using the original line. Note that the actual median is the first sample point at which the approximated area under the curve met or exceeded 0.5, hence an odd-valued result is not possible due to the sample interval of 2 ms. The actual and predicted variances agree in the first three digits for all cases, although the absolute difference is increasing.

Table 3-5 shows the results from numerical convolution for another lognormal density with values $\Delta T = 100$ and $R_C = 1$. In this case, the actual means are in general agreement with the predictions, but decrease very slightly over time, possibly the result of approximation errors. The actual medians are within 1 sample of the predictions in all instances except for the first two rows. The variances for this convolution example, however, match all predictions exactly in all 5 digits compared. These results lend good support for believing that the modified model of motion exhibits the desired time varying properties. In addition, Table 3-6 presents the logarithm of the sample at which the area under the convolution result first exceeds the area for each of several standard deviation multiples on the standard normal curve, minus the mean and divided by the multiple. In other words, the area is used to identify the point that should be *x* SD from the mean if the curve were truly normal, and the SD is then estimated through division by *x*. These values can be compared to one another, and to the expected standard deviation of the underlying normal curve, in order to assess the lognormality of the results of the convolutions. In general, these values are good, although the estimates farthest from the mean tend to be the worst. Many of these differences are larger than can be explained by the measurement error due to the sample interval. This may indicate that the tails of the convolution results are not exactly like those of a true lognormal density. Also, estimates from the right of the mean are generally higher than expected while those from the left of the mean are lower, indicating that the tails drop-off too quickly.

Table 3-4. Results of numerical convolution using the lognormal density for $\Delta T = 600$ and $R_C = 1$ generated from the revised regression line.

Conv.	Actual Mean	Pred. Mean	Actual Median	Pred. Median	Actual Var.	Pred. Var.
1	1199.99	1200	1184	1183	40701	40712
2	1799.99	1800	1784	1783	61052	61069
3	2399.99	2400	2384	2383	81403	81425

Table 3-5. Results of numerical convolution using the lognormal density for ΔT = 100 and R_C = 1 generated from the revised regression line.

Conv.	Actual Mean	Pred. Mean	Actual Median	Pred. Median	Actual Var.	Pred. Var.
5	599.98	600	582	584	20356	20356
8	899.97	900	882	884	30534	30534
11	1199.96	1200	1182	1183	40712	40712
14	1499.96	1500	1482	1483	50891	50891
17	1799.95	1800	1782	1783	61069	61069
20	2099.94	2100	2082	2083	71247	71247
23	2399.93	2400	2382	2383	81425	81425

Finally, although the revised regression line for the inverse variances did not drastically alter the predicted means and variances for the data sets, K-S statistics were calculated for all data sets using the revised estimates of the underlying normal curves. All of these values were comparable to those resulting from the original experiment (nearly equal in many cases), except for the data set for an elapsed time of 600 ms. The K-S statistic for this set jumped from .0854 to .0961. Both the predicted mean and variance of the underlying normal distribution for this data set changed more significantly than the corresponding predictions for any other data set. As with the K-S statistic for the original predicted normal curves, the

Table 3-6. SD estimates from results from numerical convolution using the lognormal density for ΔT = 100 and R_C = 1 generated from the revised regression line.

Conv.	Pred. SD	SD at -3 SD	SD at -2 SD	SD at -1 SD	SD at +1 SD	SD at +2 SD	SD at +3 SD
5	0.2345	0.2299	0.2292	0.2309	0.2348	0.2376	0.2423
8	0.1924	0.1892	0.1893	0.1892	0.1932	0.1948	0.1981
11	0.1670	0.1643	0.1637	0.1652	0.1678	0.1686	0.1716
14	0.1496	0.1475	0.1472	0.1481	0.1501	0.1509	0.1537
17	0.1366	0.1346	0.1344	0.1356	0.1372	0.1380	0.1402
20	0.1266	0.1252	0.1249	0.1257	0.1270	0.1278	0.1297
23	0.1185	0.1172	0.1173	0.1175	0.1186	0.1193	0.1211

K-S statistic for the revised curves was larger for data sets whose actual means varied most from the predicted means.

Figure 3-4 shows the fitted normal curves using the revised regression model. Data sets for elapsed times from 600 ms to 1800 ms are presented. The curve for 1500 ms produces the largest value of the K-S statistic when compared to the respective data set. Note that the curves producing the larger values of the K-S statistic (including the curve for 1500 ms) contain more area than the respective data sets just to the left of the peak. This missing area in the data sets appears to be shifted closer to the peaks than is expected for the fitted curves. Once again, this effect is believed to result from bias introduced when generating the data sets.

It is desirable that a model of a performer's motion through the score should provide consistent position estimates when the elapsed time between observations can vary. Examining the initial model of motion has led to a revised model of motion that exhibits such consistent behavior. Based on the calculated K-S statistics, the revised model accounts for the data nearly as well as the original model. The revised model determines the inverse variance of the rate ratio from the elapsed time as follows:

$$\frac{1}{\beta^2} = 0.02948 \times \Delta T \qquad [3.6]$$

This estimate can be converted into an estimate of the variance of the underlying normal distribution, σ^2, using equation 3.5. This variance, along with an estimated mean of $\mu = \ln R_C - \frac{1}{2}\sigma^2 + \ln \Delta T$, can be substituted into equation 3.2 to provide a distance density for the score-following model. The next section will provide a more comprehensive overview of implementing this model of a performer's motion as part of both the stochastic score-following model and a complete automated accompaniment system.

3.5 Implementing the Model of Motion

The formal model of a performer's motion through the score is based upon the beliefs that the rate is always positive, the actual average rate in the recent past will be a good estimate of the average rate over the immediate future, and the majority of changes to the rate can be treated as "small" and "random" (with "large" changes being infrequent and either explicit in the score or self-cancelling over time). These beliefs support the use of a lognormal density for modeling a performer's rate and the score distance traversed over a fixed period of time. Examination of actual performances has given reason to be cautiously optimistic regarding the use of such a model for automated accompaniment. Examination of numerical convolution of lognormal densities, in combination with performance data, has also indicated that lognormal convolution is at least a consistent model of subsequent location of a performer given an initial starting point, an average rate from the recent past, and an elapsed time less than 3 seconds.

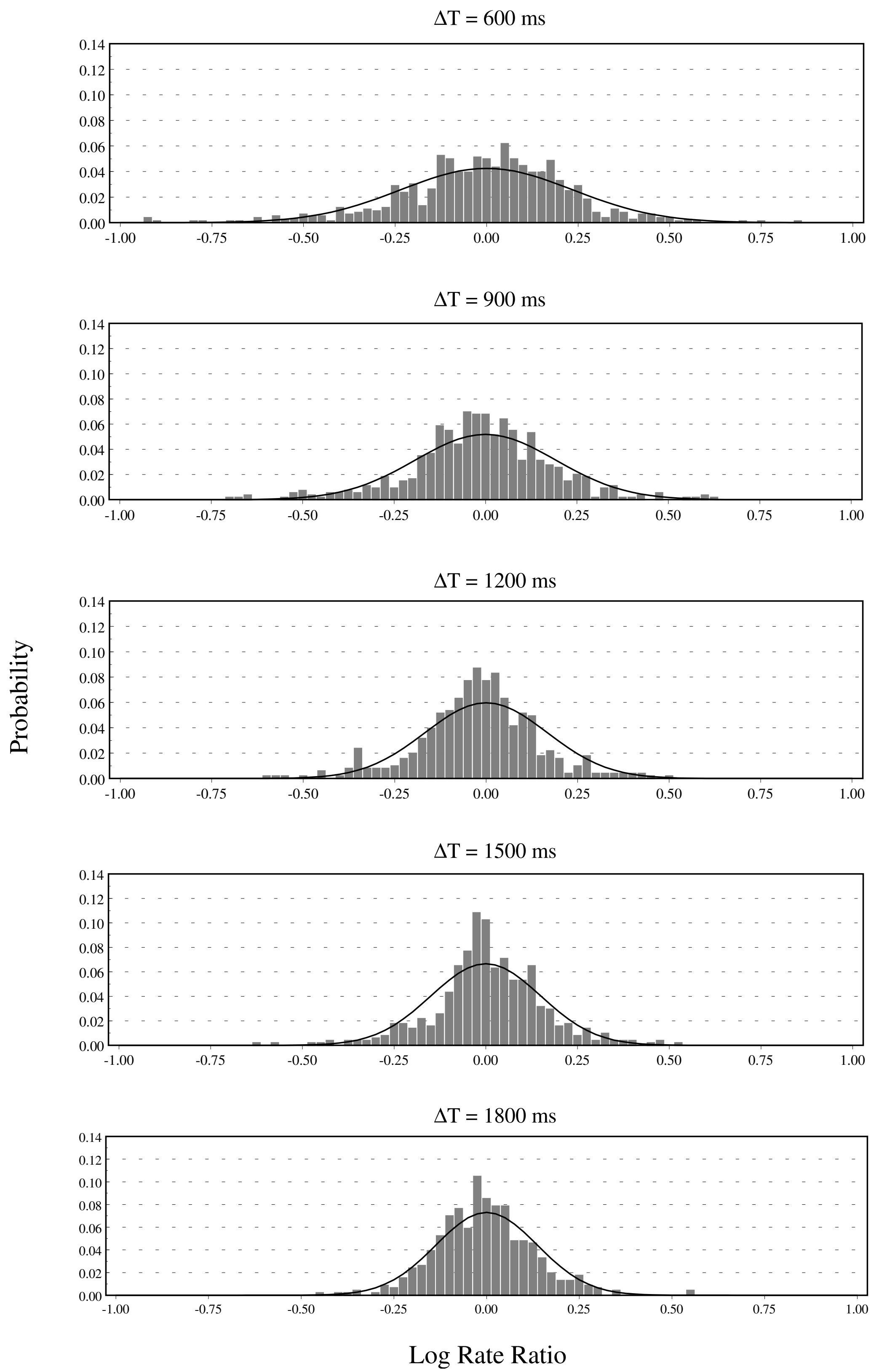


Figure 3-4. Normal curves fit to rate ratio data sets using the revised regression line for the inverse variances.

Development of the model of motion has to this point not addressed several concerns relating to implementation of a real-time accompaniment system. These concerns include both real-time execution and error propagation as discussed in the previous chapter, as well as fulfilling all the assumptions of the model of motion. More specifically, these issues include:

1. Selecting a sample interval for numerically representing the distance density, which in turn requires specifying a minimum for the elapsed time.

2. Efficiently generating samples from the parametric distance density, relative to computing the convolution via the discrete Fourier transform.

3. Generating initial and subsequent rate estimates in a manner consistent with the model of motion.

4. Extracting event lengths from a symbolic score in a manner consistent with the model of motion.

This section describes specifics of system implementation that address each of these concerns and can satisfy the requirements of using the score-following model as part of a fully functional automated accompaniment system.

The error analysis in Chapter 2 indicated that the error from approximating the convolution integral has the potential to propagate over time. A method of bounding this error for a maximum number of model updates also was provided. For this bound to be useful, an error bound must be given for approximating the integral of the product of the functions participating in the convolution. We now know that the distance density participating in the convolution is a lognormal density function, but the density describing the performer's source position is not explicitly known. However, most often this latter density will result from several successive convolutions by a lognormal density and multiplications by the observation density (as well as normalization). Since the observation density (as specified in Chapter 2) will be constant over the length of an individual event; it is reasonable to expect that, within an event, the source position density will often be as smooth or smoother than the lognormal density of minimum variance (*i.e.*, the maximums of its derivatives will be as small or smaller). Thus, examining the relative error from approximating the self-convolution of the sharpest lognormal density (the one with minimum variance) can provide a reasonable estimate of an upper bound of the convolution error in the score-following model.

The sharpest lognormal density in the model of motion will be the one with the shortest elapsed time. For score following, the minimum elapsed time between observations will be selected at 100 ms. Thus, the accompaniment system will be able to adjust its position and tempo at a rate of 10 Hz. Likely this response rate is comparable to human response time and is large enough to allow time for both model updates over a reasonably large score window and the signal processing required to generate observations. Given this minimum elapsed time and the model of motion presented in the last section, self-convolution of a lognormal density with $\mu = -0.1461$ and $\sigma^2 = .2921$ should provide a good estimate of the maximum relative error from a single convolution.

Recall that a convolution integral effectively flips one function prior to multiplication. Since a lognormal density is nonzero only over the positive part of the number line, flipping one of the densities will generally leave only a limited range where one of the functions is nonzero. For each point in the result of convolution, numerical integration needs to consider only this region. Table 3-7 shows the results of numerical integration applied to self-convolution of the lognormal density for $\Delta T = 100$. Values are given for different sized nonzero regions sampled at different intervals. Regions of size 86, 255, and 437 ms were selected. These points on the lognormal density correspond to the integer nearest the mean, +2 SD, and +3 SD points on the underlying normal density.

For purposes of bounding error propagation, we are concerned with the relative error between the approximated and actual values of the density function. In particular, the maximum relative error over all points calculated for the density function is of interest. In Chapter 2 it was shown that to guarantee a maximum relative error of 2 after 5000 applications of the score-following model, the relative error for a single convolution must be within a range of .99993 to 1.00007. Bold entries in the table indicate the first sample interval at which the relative error of an estimate (compared against the Simpson approximation using the most samples) first falls within this range. Note that as expected, Simpson's method generally achieves this bound with about 33 samples.

Interestingly, midpoint approximation achieves the bound with the same or fewer samples. This self-convolution type of example turns out to be an especially good instance for midpoint approximation, but this is not in general the case. Table 3-8 shows results of numerical integration over a portion of just the lognormal density for $\Delta T = 100$. In this case, Simpson's rule converges more rapidly and still achieves the target error bound using fewer than 33 samples.

However, in practice it would seem that either midpoint approximation or Simpson's approximation could be used to satisfy the error bounds for the score-following model. In the examples considered, sufficient relative errors are achieved using a sample interval between 5 and 13 ms. Approximating area in smaller regions would likely have a larger relative error using sample intervals in this range. However, such points in the result of the convolution will be small in value. Thus a sample interval of 10 or 12 ms would seem sufficient to limit the propagation of error.

Table 3-7. Results of numerical self-convolution of the lognormal density for ΔT=100 and R_C=1. Relative errors compare each value against the approximation using Simpson's rule with 513 samples. Results in bold indicate the first sample interval yielding a relative error between 0.99993 and 1.00007.

Nonzero Region	No. of Samples	Sample Interval	Midpoint Approx.	Relative Error	Simpson's Approx.	Relative Error
0-86	5	21.5	0.00175930149	1.00902514776	0.00150642839	0.86399297865
	9	10.75	**0.00174361993**	**1.00003118950**	0.00174877654	1.00298869753
	17	5.375	0.00174356410	0.99999916524	**0.00174358464**	**1.00001095047**
	33	2.6875	0.00174356555	1.00000000190	0.00174356506	0.99999972048
	65	1.34375	0.00174356555	1.00000000000	0.00174356555	1.00000000064
	129	< 1.0	0.00174356555	1.00000000000	0.00174356555	1.00000000000
	257	< 1.0	0.00174356555	1.00000000000	0.00174356555	1.00000000000
	513	< 1.0	0.00174356555	1.00000000000	0.00174356555	1.00000000000
0-255	5	63.75	0.00273973160	0.96843384454	0.00304972192	1.07800848939
	9	31.875	0.00282593900	0.99890623448	0.00280103024	0.99010154371
	17	15.9375	0.00282926130	1.00008059339	0.00282785252	0.99958262075
	33	7.96875	**0.00282903130**	**0.99999929390**	**0.00282911063**	**1.00002733356**
	65	3.984375	0.00282903330	1.00000000123	0.00282903263	0.99999976381
	129	1.9921875	0.00282903330	1.00000000000	0.00282903330	1.00000000041
	257	< 1.0	0.00282903330	1.00000000000	0.00282903330	1.00000000000
	513	< 1.0	0.00282903330	1.00000000000	0.00282903330	1.00000000000
0-437	5	109.25	0.00022695445	1.05253477457	0.00023630446	1.09589683750
	9	54.625	0.00021276661	0.98673656131	0.00022129926	1.02630800839
	17	27.3125	**0.00021562739**	**1.00000385775**	0.00021466829	0.99555591598
	33	13.65625	0.00021562992	1.00001558855	**0.00021562464**	**0.99999110293**
	65	6.828125	0.00021562652	0.99999984228	0.00021562770	1.00000530087
	129	3.4140625	0.00021562656	1.00000000034	0.00021562655	0.99999994720
	257	1.70703125	0.00021562656	1.00000000000	0.00021562656	1.00000000012
	513	< 1.0	0.00021562656	1.00000000000	0.00021562656	1.00000000000

Table 3-8. Results of numerical integration of the lognormal density for ΔT = 100 and R_c = 1. Relative errors compare each value against the best approximation using Simpson's rule. Results in bold indicate the first sample interval yielding a relative error between 0.99993 and 1.00007.

Nonzero Region	No. of Samples	Sample Interval	Midpoint Approx.	Relative Error	Simpson's Approx.	Relative Error
0-86	5	21.5	0.49903584910	1.00512358429	0.48765203261	0.98219508635
	9	10.75	0.49697316649	1.00096907125	0.49670389444	1.00042672207
	17	5.375	0.49661031580	1.00023824242	**0.49649475003**	**1.00000547782**
	33	2.6875	**0.49652158653**	**1.00005953005**	0.49649205780	1.00000005532
	65	1.34375	0.49649941699	1.00001487772	0.49649203395	1.00000000729
	129	< 1.0	0.49649387685	1.00000371913	0.49649203055	1.00000000045
	257	< 1.0	0.49649249195	1.00000092976	0.49649203035	1.00000000003
	513	< 1.0	0.49649214574	1.00000023244	0.49649203033	1.00000000000

In addition, windowing the score position density function will require handling long rests as special cases. For example, it is not uncommon for vocal performances to contain periods of 10 or more seconds where the vocalist does not sing. Fortunately, long silences can be reliably recognized. Given a threshold for the duration of such silences, and assuming the position estimate prior to the silence is reasonable, the tracking system can be forced to skip past the next expected large rest when a long silence is identified. The density function's window can be centered around the score just beyond the rest. The score position density can be initialized to expect the next observation to be the performer beginning that section of score. Since a significant rest will likely occur at least once per piece, the actual period of time over which errors can propagate will be less than half the length of a song—certainly less than 4 minutes. Under this scenario, a sample interval of 10 or 12 ms should be more than sufficient.

Having determined an appropriate sample interval, the next concern for implementing the model of motion is efficiently sampling a lognormal density. For a window encompassing N samples, this sampling operation requires at most N evaluations of the lognormal density:

$$\frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

each of which requires calculation of a natural logarithm and an exponentiation. Enumerating a set of allowable values for μ and σ would enable use of a lookup table that could be precomputed, but would also introduce another possible source of error in the score-following model. Fortunately, for a constant sized window and sample interval of ΔS, the sample locations can always be specified as:

$$x = n \cdot \Delta S, \quad n = \left(-\tfrac{N}{2} + 1\right), \cdots, \left(\tfrac{N}{2}\right)$$

Since the values of ΔS and N will remain constant during a performance, the natural logarithm may be precomputed over the range of x. Using this precomputed logarithm table, N samples of the lognormal density are then calculated in real time during an actual performance, allowing σ and μ to change for each application of the model. The logarithm and exponential are the most costly calculations for each sample, so precomputing the logarithm nearly halves the computation time for sampling the density. For any reasonably large window size, computing the lognormal density function for each model update requires trivial computation time compared to the cost of computing convolution via the FFT.

Another implementation issue is generating a rate estimate during an actual performance. The density functions in the model of motion were based upon predicted rates calculated over consecutive notes performed within a time period of 2 to 5 seconds. Also, the boundaries of this time period always corresponded to the onset of a note. An accompaniment system can produce a similar rate estimate during a performance by maintaining a sufficiently long buffer of time-stamped, point position estimates according to the following policies:

1. A position estimate (*e.g.*, the maximum likelihood estimate from the score position density) is placed in the buffer only when a new position estimate crosses a note onset, as indicated by the score, relative to the previous position estimate.

2. Rate estimates are calculated using the most recent buffer entries spanning the required elapsed time. A reasonable policy might be the first set of buffer entries spanning at least 3 seconds.

3. The buffer is emptied whenever an estimated position precedes any estimate already in the buffer (*i.e.*, the system thinks the performer has backed up) or exceeds a reasonable maximum distance from the previous estimate (*e.g.*, a score distance requiring the performer to have increased the rate by a factor of 2 over the previous rate estimate).

4. In situations where the buffer does not contain sufficient position entries to provide a rate estimate, the tracking system can use either the last provided estimate or the default tempo provided in the score. In practice, it is likely that the performer may wish to provide a good default tempo prior to starting the performance.

Note that these rate estimation policies attempt to correct or second-guess values calculated from the point position estimates only in extreme cases. Likewise, the model of motion has been specified using only observed errors in the performance and does not attempt to model rate errors that result from poor position estimation. Generally this approach is reasonable because if the model of motion and the observation densities together cannot be used to estimate position and rate with variance at least as small as those densities fitted to data in this section, then it is unlikely that the proposed stochastic tracking system will be useful.

Finally, the model of motion depends upon knowing relative lengths of the events extracted from a symbolic score. To some extent, specification of these lengths will depend upon the specific definition of the events to be extracted. At a minimum, however, event lengths will depend upon relative note durations (*i.e.*, rhythm) and relative changes in tempo, both between sections of the score and within individual phrases. Rhythms can be translated literally, in sections where the tempo is not indicated to change noticeably, by using a reasonable estimate of the tempo. The latter may be obtained either from tempo or metronome markings in the score, or from initial estimates given by the performer. Similarly, this tempo estimate can be changed as explicitly indicated in various sections of the score.

The only tricky estimation of event length will occur when markings indicate a gradual as opposed to a sudden change of tempo, or when the performer introduces sudden tempo changes not explicitly marked. In the latter case, the performer will want to indicate such changes explicitly if problems arise during an initial rehearsal. In the former case, a slight, proportionate lengthening of each subsequent note within the marked region may prove sufficient to keep the variance of the motion in that region similar to the proposed model. A simple solution would be to estimate a tempo at the beginning and end of the marked region, assume a linear decrease or increase over that region, and adjust the event lengths accordingly. In addition, since only extreme changes in tempo were excluded from the modeled data, mild changes (such as a slight slowing at the ends of phrases) may not require adjustments to the event length estimates.

In summary, the model of motion described in this chapter can be implemented as follows:

1. Event lengths are extracted from a symbolic score based upon a combination of relative note durations (rhythm); tempo estimates derived from defaults, the score, the performer, or a human accompanist; and possibly predetermined knowledge specific to particular kinds of events.

2. Score position is estimated once every 100 ms when observations are available. Thus, the accompaniment system can respond to the performer at a rate of about 10 Hz.

3. Rate estimates are obtained from a buffer of the most recent score position estimates, according to the policies previously described.

4. For each update of the score position density function, equations 3.6 and 3.5 are used to convert an elapsed time into the parameter σ for a lognormal density function describing the performer's motion for an estimated rate and elapsed time. This latter density is specified by equation 3.2. The parameters can be generated as needed, and a numerical representation of the corresponding lognormal density can be computed in real time.

5. The score position density and distance density are represented numerically with a sample interval between 10 and 12 ms. Intervals of this size should satisfy desired error bounds for a minimum elapsed time of 100 ms between updates of the score position density function.

6. Long periods of silence can be identified accurately and reliably. Consequently, they can be used to recognize the performance of long rests. Before updating the score position density based upon a successive observation, the density function must be altered to indicate a strong likelihood that the performer was at the end of the long rest just prior to the new observation.

The model of motion is implemented in this manner for subsequent construction and evaluation of an actual automated accompaniment system.

3.6 Comments on Assumptions of the Score-following Model

The final score-following model presented in Chapter 2 is based upon several simplifying assumptions. The convolution assumption used to expedite computation is most obviously relevant to the model of motion presented in this chapter. Examination of actual vocal performances has shown that a model based upon convolution of lognormal densities is reasonable with respect to at least mean and variance of a performer's rate over short but useful periods of time (*e.g.*, less than 6 or 8 seconds), and is not a bad model with respect to empirical distributions derived from data with some likely biases. The theoretical glue holding this model together relies upon two important additional assumptions about vocal performances. First, over the relatively short periods of time examined, the performer attempts to maintain an essentially constant average rate. Second, variations of the average rate over subintervals of such time periods can be viewed as essentially random and proportionate to the targeted average rate. Such a model may be very useful for score following. If the essentially linear propagation of mean and variance is good for short periods of time, then frequently occurring, discriminating observations may be able to effectively "tug" the position density function as it moves slightly off-track from time to time. Finally, use of a lognormal density satisfies the requirement of an exclusively positive-valued rate. Since the distance density is therefore nonzero only in the positive direction, the model avoids any problems with making a convolution assumption when the performer is near the beginning of the score.

The support for an assumption of constant average rate over short time periods also ameliorates concerns about using a single value for the estimated distance. While such an approach may miss subtle position changes in the extreme short term, the convolution process indicates that these variations are at least unlikely to be significant over time. Substituting f_{Prior} for $f_{J|D}$ in the final score-following model similarly becomes less of a concern, particularly since rate is likely to be estimated by using regions of high density in recent estimates of f_{Prior}.

The model of motion presented here may appear similar to a variety of techniques and approaches familiar to different readers. For now, rather than enumerating these related viewpoints and explicitly comparing them against the model of motion, the following important aspects and advantages to the presented model are simply listed without further discussion:

1. In this model, the score position of a performer is real-valued, and the statistical description of how it changes over time is continuous. By applying familiar forms of numerical analysis, discretization of the model can achieve desired precision and accuracy while minimizing computational cost.

2. Rate and time are treated explicitly, and are permitted to assume values within a continuous range.

3. Internally, the model is consistent as elapsed time between updates varies. As long as observations are generated in a consistent and time-invariant manner, this property allows the score-following model to provide consistent position estimates even when the elapsed time between observations may vary. The reliability of such estimates will of course depend upon the relevance of the omitted observations or the unexamined signal.

4. Assumptions needed to support the model are fairly general and are likely to be easily and frequently satisfied. To the extent that unbiased data can be readily obtained, appropriate density functions (possibly other than the lognormal density) can be fit to actual, observed motion in the situations of interest.

5. In certain cases, possibly depending upon the density functions used in the model, certain well-understood methods of functional analysis, statistical analysis, and statistical modeling can be applied to assess the model and indicate possible ways for improving it.

A consideration of specific, related approaches is undertaken in a subsequent chapter.

Finally, the following aspects of the model are deserving of further investigation and possible enhancement:

1. Data obtained from within a single performance may exhibit biases due to high-level structure in music. It would be useful to obtain performance data with less possibility for biases. The given model could be fit to that data for purposes of comparing it to the results already obtained.

2. Other density functions can be shown to behave as the lognormal does under convolution. For instance, a gamma density also describes a positive variate whose mean and variance are additive under convolution; and it has been formally shown that the lognormal density is a *generalized gamma convolution* (Thorin 1977), implying that it is well approximated by a sum of independent gamma variates. Such alternative density functions may provide better fits to observed data while maintaining all the desirable properties of the lognormal distribution.

3. The present model of motion does not explicitly account for expected changes in rate. It would be useful to extend the model to account for expected rate changes. Such an enhanced model might permit a more straightforward derivation of event lengths from a score. A model based on a constant acceleration might be possible.

4. Beyond the use of estimated rate, the present model makes no further attempt to account for correlation between current observed rate and successive actual rate. It would be useful to look for additional features or patterns in the score, or other observable behavior, that might be strongly correlated with the motion of a performer.

5. The present model is generalized across a number of performers, pieces, and styles. It would be useful to investigate possible methods for adapting the model to a particular performer, piece, or style. Such adaptation might be based on additional background knowledge, and might be executed in real time during or after each performance.

As always, additional understanding of relevant variables and their interactions can lead to enhanced models of potentially greater accuracy and predictive value.

The model of motion as presented in this chapter was employed as part of the score-following model within an automated accompaniment system. The immediately following chapters present definitions of various events and the corresponding observation densities. Various combinations of these statistical models were used to accompany recorded performances, and subsequently live performances as well.

Chapter 4

Events Based on Observing Fundamental Pitch

4.1 Defining Events and Modeling Observations

A stochastic model for tracking vocal performances (or musical performances in general) was presented in Chapter 2. The model provides a method for calculating a density function over the score. The probability that the singer is within any region of the score can be calculated from the density function. The equations of the model incorporate three distinct pieces of information—a statistical description of the previous location of the performer, an estimate of the amount of score performed during the elapsed time, and a recent observation of the singer. Application of the statistical model requires estimation of two density functions appearing in the equations. One function specifies the likelihood that the singer has actually performed a certain amount of score given an estimate of the amount performed. Chapter 3 presented a definition of this density function based on analysis of actual vocal performances. The second function specifies the likelihood of making a particular observation given the current score position of the performer. This chapter addresses estimating such an observation density when the observation consists of fundamental pitch automatically extracted from a singer's performance.

Before discussing fundamental pitch and methods for its detection, it is important to carefully describe what constitutes an observation. It is assumed that observations are values that result from repeatable processing of signals from the environment. The type of the observation is characterized by the process used to generate the observation. For instance, identification of fundamental pitch would yield one type of observation, while detection of maximum amplitude would yield another type. Observations are likely to be reported periodically, the result of processing successive, small portions of a signal. Different types of observations may be reported for the same portion of a signal, but no portion of a signal is associated with more than one value for a given type of observation. The mapping from a portion of a signal to an observation value is a true function. However, the value reported for a given portion of a signal may be influenced by the processing of preceding portions of that signal. Thus techniques like filtering or smoothing may be used to extract observations. The possible range of values for an observation will contain at least two distinct values and may or may not be continuous. This insures that an observation has at least some potential to discriminate between different events or score positions, but permits for a useful level of quantization in reported observation values.

The purpose of observations is to help the tracking system discriminate score position, or at least distinguish among events in the score. Recall that a score was defined as a sequence of events, an event having both an associated length (relative duration) and an observation distribution. The observation distributions indicate the likelihood of observing any possible value for a given type of observation when the singer is performing a given event. Events essentially provide a way of simplifying the description and modeling of observation distributions. Ideally, one would like to specify a highly accurate observation distribution for every point in the score. This approach would enable modeling of continuous and subtle changes in likelihood for the value of a given type of observation. In reality, however, it will not be feasible to discern or to estimate changes in observation likelihood at such a fine resolution. Estimating observation distributions for individual events is one alternative. This approach requires only a single distribution for each of several, nonoverlapping regions of score. Mathematically, this corresponds to approximating the observation density $f_{V|I}(v \mid i)$ by a function specifying probability density for an observation conditioned on the event, say $f_{V|E}(v \mid e)$. Note that this change amounts to a substitution of the conditioning variable. When it is possible to identify many score regions that span points with very similar distributions for a given observation type, this approximation can be quite good.

Specifying observation distributions per event rather than per score position is a useful simplification. In reality, however, it is still not realistic to explicitly define observation distributions specific to every event in a given score, prior to a performance. This approach would require nontrivial amounts of knowledge about performances of a particular score, and perhaps about a particular singer's rendering of that score. While an expert in score following, along with a willing singer and accompanist, might be able to construct such a system by examining several performances, the result would not be a general purpose tracking system. Consequently, it is necessary to identify important pieces of information, or factors, that significantly influence the value of a given observation type. Factors will include information in the score. For instance, the scored pitch for an event will influence the pitch detected during performance of that event. Factors may also include information provided by the singer prior to a performance, or other information that can be detected from the environment.

The use of factors mathematically corresponds to approximating the observation density $f_{V|E}(v|e)$ by a function specifying probability density for an observation conditioned on the values of important factors, say $f_{V|A[1],A[2],...,A[\mathrm{N}]}(v \mid a[1],a[2],...,a[\mathrm{N}])$. This approximation may be very reasonable providing all important factors can be identified, the values of the factors are known for each event, and sufficient data can be collected to accurately estimate the distribution over the space defined by the factors' possible values. Note that a similar substitution was made in Chapter 3 when tempo and elapsed time were substituted for estimated distance in the density function characterizing actual score performed. Observation densities estimated in this chapter and subsequent chapters will result from applying this approach.

An ideal type of observation will offer high discriminatory ability while permitting for easy specification of highly accurate distribution estimates. Ease of accurate estimation depends on several considerations. As already mentioned, either it must be feasible to define the observation distributions for every point in the score, or the observation distributions for many contiguous points must be similar enough to allow definition of events. The former case requires some nice property that enables parametric specification of a closed-form, continuous or piecewise continuous density function over score position and observation value. The latter case requires that the observation distribution for an individual event should not result in poor approximations of the distributions for individual points within that event. Additionally, all factors that influence the observation distribution for a given event or position must be known. Distributions developed in ignorance of important factors may not adequately approximate the actual distributions for an event or position. This requirement is identical to saying that all relevant conditioning variables and their values for each event must be known. Finally, the observation densities must characterize the generation of observations in the real world. Thus, it must be feasible to collect and process sufficient data to precisely define the observation density functions. It is not sufficient just to know the general shape or family of the functions—the numerical values at points in the functions must be known.

Since the observation distributions will be incorporated as part of the stochastic score-following model, it is important that the sensors designed to report selected types of observations do not violate assumptions of the model. Specifically, the value of each observation must be independent of both the previous position of the performer (source position) and the estimated amount of score performed (*i.e.*, independent of tempo and elapsed time since the previous observation). Thus, for each reported observation, signal processing is applied only to a limited amount of the performance. Implementation of the model further requires that each observation is reported as a fixed value, not as a statistical distribution over possible values. Consequently, the applied signal processing must generate a single estimate for each observation—methods that produce a distribution over several possible values are not useful to the score-following model. These requirements will constrain the applicable methods for generating observations of a given type. Finally, time varying properties of the model considered in Chapter 3 led to a question regarding the duration of signal used for generating each observation. Specifically, does extracting observations from signal regions with varying durations yield different score position estimates (assuming the observation distributions consider this duration as a factor) than consistently extracting observations from signal regions with identical durations. The answer to this question does not impose a direct constraint on defining observations, but would indicate how the frequency with which observations are reported impacts accuracy of position estimation. Since the work presented in this document constitutes an initial investigation of the stochastic score-tracking model, each observation of a specific type will be extracted only from signal regions with identical or nearly identical durations. Subsequent work discussed in this document does not explicitly address the question raised in Chapter 3, leaving it open for future investigation.

In summary, there are five general requirements that an ideal type of observation will satisfy completely. The utility of a given type of observation can be judged with respect to how well it fulfills these requirements. The requirements are as follows:

1. **Position Discrimination:** Observations must discriminate between positions (or at least events).

2. **Ease of Specification:** Observation distributions can feasibly be specified for every point in the score. Alternatively, observation distributions can be specified for clearly defined regions of the score that will constitute events, providing that the distribution for each event adequately approximates the distributions for each point within that event.

3. **Availability of Significant Factors:** All conditioning variables necessary to accurately approximate $f_{V|I}(v \mid i)$ for each score position or event are known. The values these variables assume during each event are also known.

4. **Ease of Estimation:** It is tractable both to collect and to process sufficient data to derive accurate and precise estimates of $f_{V|I}(v \mid i)$ over the space of conditioning variables.

5. **Model Consistency:** Observations must not violate assumptions of the stochastic score-tracking model. Specifically, they must be independent of the previous position of the performer and the estimated amount of performed score (*i.e.*, tempo and elapsed time between observations). They must also be obtained as fixed values in real time.

These criteria will be important when defining observations for use in tracking vocal performances.

4.2 Fundamental Pitch in Vocal Performances

While criteria for an ideal type of observation have been provided, realistic types of observations inevitably fail to satisfy these requirements completely. Fundamental pitch of vocal performances is no exception. At a high level of abstraction, fundamental pitch can be viewed as the key on the piano that, according to human perception, essentially replicates or is analogous to a given tone produced by a singer or instrument. Considering even this abstract and imprecise definition allows us to recognize several shortcomings of using fundamental pitch as an observation.

Initially fundamental pitch seems like a reasonable choice for a type of observation. For instance, it is explicitly notated in musical scores and competent singers generally expend great effort to control it. As indicated in Chapter 1, however, it is common to find sections of a score where two or more successive notes have the same fundamental pitch. It is not rare to find extended sections of five, ten, or more successive notes with the same fundamental pitch. This knowledge of scores warns that fundamental pitch may not always provide adequate position discrimination, the first criterion for an ideal observation type. Some regions of the score are likely to be indistinguishable based on fundamental pitch. Occasionally these regions may encompass many contiguous notes, preventing the estimated score distance discussed in Chapter 3 from helping to disambiguate position if a singer alters the tempo significantly within these regions.

Notes in a musical score specify fundamental pitch. Thus, for purposes of observation distributions based on detection of fundamental pitch, it would seem reasonable to establish a one-to-one correspondence between notes and events. Ignoring certain factors such as the individual singer and the specific score, it might be feasible to provide an observation distribution for each such event based only on local factors in the score. However, it is not so clear that each such distribution would adequately approximate the distribution at every point within the respective event. One problem may be the vibrato technique commonly applied by singers. This technique results in the sung pitch intentionally oscillating around the scored pitch. Properties of vibrato might cause different regions of a note to have different distributions for observed fundamental pitch. Also, many notes in a score may not be sung on a single vowel, but may be sung on a syllable with a *diphthong* (one vowel changing to another) and possibly consonants. Distributions over observed fundamental pitch might be affected by the different vowels and consonants, or by the transition from one phone to the next. These considerations, and probably several others, make it unlikely that distributions for observations based on fundamental pitch are easily specified.

Many factors influence the rendering of a musical score and the production of fundamental pitch. In addition to vibrato, there are numerous ways that singers alter pitch for expressive purposes. Often these changes are not explicitly notated in the score. *Portamento* is a technique where the pitch of one note is made to change continuously, or "slide", into the pitch of the next note. *Ornamentation* is when the singer intentionally adds discrete pitches to the score, shortening the scored notes to make room for the additions. This technique is applied to embellish or decorate the melody, especially when repeating a section in pieces written in certain styles. The placement and nature of these intentional changes often depends upon higher level structures in a musical score, such as phrases, harmony, and cadences. Besides intentional changes, singers sometimes make subconscious or accidental alterations to pitch. During memorization of a piece, they may alter a pitch so that it differs from the score but is not harmonically offensive. When singers are tired or not in top health, pitches at the extremes of their ranges may be rendered flat or sharp. The list of possible alterations to pitch is extensive, and no enumeration of all the factors influencing these changes is readily available. Even if such knowledge could be obtained, it is unlikely that all important factors could be automatically measured or made directly available to a computer tracking system. These considerations limit the availability of many significant factors that influence the observations based on fundamental pitch.

Collecting sufficient examples of vocal performances is not an easy task. Commercial recordings are not usable. They seldom contain a singer recorded in isolation, and reliable sound source separation (to extract the voice sound from the rest of the recording) is an unsolved problem. Recording singers is a time consuming task. Proper recording requires at least three, often difficult to schedule experts—a singer, an accompanist, and an audio technician. Also, the musicians must have practiced the pieces to be recorded, at least on an individual basis if not as an ensemble. If willing volunteers can not be found, collecting large numbers of recordings can require nontrivial financial resources. However, data collection was done for this project. As described in Chapter 3, twenty examples of different performances were recorded for analysis purposes. These examples are by no means comprehensive. The examples do span all primary voice parts, provide ample pitch variety within the ranges of the singers, and include a variety of pieces in different languages, styles, and genres. Hopefully, the amount of data collected for this study can provide reasonable estimates of distributions based on only a few conditioning variables. Even if this were not the case, one can at least feel confident that subsequent collection of additional data, though time consuming, could eventually correct any problems. Thus, for a very limited number of conditioning variables (as will be applied in this study), it is feasible to estimate distributions for observations based on fundamental pitch. However, if it is essential to use many more conditioning variables to adequately approximate the observation density, estimation may not be feasible.

Finally, observations based on fundamental pitch may violate assumptions of the score-following model. Now fundamental pitch can in fact be reported as a fixed value, and the processing required can be done in real time. A software implementation that achieves these objectives is described in the next section. However, as will become apparent in the next section, it is not feasible to obtain estimates of fundamental pitch that are entirely independent of the previous position of the performer and the estimated amount of score performed. Specifically, since the amount of signal processed will span a fixed duration, changes in a singer's tempo can alter what portion of the performance is processed when detecting fundamental pitch. Thus the observation distribution may differ depending on the actual tempo, especially for score positions near the beginning of notes where the processed region of signal may cross a note transition. For faster tempi, more signal from the previous note will appear in the processed region. Actual tempo, of course, significantly influences both estimated tempo and previous position.

Many of the criteria for an ideal observation are difficult to achieve. However, it is important to put these criteria into proper perspective. Failure to satisfy these requirements completely was expected and does not imply that building a general purpose vocal tracking system is impossible. The thesis that considering multiple observations not only helps vocal tracking but is actually necessary for robust tracking was presented in Chapter 1. If fundamental pitch could completely discriminate score position, this thesis would be wrong. The remaining criteria present requirements for achieving perfect accuracy of observation density estimates. In reality, however, reasonable approximations may be sufficient for adequate and robust tracking. With respect to simplifying distribution estimation and maintaining model consistency, arguments supporting approximations have already been provided. Support for approximations that simplify specification and reduce the number of considered factors will require empirical investigation. However, it is safe to say that even the best human accompanists do not make use of or even know all the factors that influence fundamental pitch, nor can they detect arbitrarily small differences in position based purely on changes in pitch. Complete accuracy of observation distributions may not be worth the effort.

4.3 Methods of Pitch Detection

From a physical perspective, sound equates to variations in air pressure. The distribution of air pressures that travels away from a sound source is referred to as a *sound wave*. Fourier's theory states that any wave (or any continuous function over the reals) can be decomposed into a linear combination of sine waves. The Fourier transform first mentioned in Chapter 2 is in fact an analytical method for expressing a given function as a summation of sine functions with different periods, each sine function possibly having a distinct scaling (or amplitude) and translation (or phase). Note that for an arbitrary function, an infinite number of sine waves may be required to represent the function with complete accuracy. A function specifying the amplitude and phase for the sine waves that compose a given sound is referred to as the *complex spectrum*, or simply the *spectrum*, of the sound.

Pitched sounds generally correspond to pressure waves that exhibit a single, repeated pattern of pressure variation. These sound waves are said to be *periodic*. The spectrum of a truly periodic sound consists only of sine waves whose periods are multiples of a single value. This value, the period of the first sine wave in the spectrum, is called the *fundamental frequency*. The sounds produced by playing notes on musical instruments are often described as periodic. In reality, these sounds are not truly periodic, but their spectra may exhibit sharp peaks spaced by a nearly constant value. The spectra are said to contain a fundamental frequency and many *harmonics*. Sine waves in the spectrum whose periods fall between the periods of the harmonics are referred to as the *inharmonics*. Sung vowels exhibit a spectral characteristic similar to sounds produced by musical instruments, although the inharmonics often have greater amplitude in sounds produced by singing than in sounds produced by many instruments.

Figure 4-1 shows a graph of the sound wave from a sung vowel and the amplitude values from its associated spectrum. A waveform as shown in the left graph displays change in amplitude over time for a fixed point in space (*i.e.*, the location of the microphone). Amplitude is a relative value indicating changes in air pressure. The line in the graph indicates an amplitude of zero, or no change from the reference. Positive values indicate an increase in pressure from the reference, while negative values indicate a decrease. In the spectral graph on the right, frequency indicates the periodicity of the sine wave. Frequency is usually measured in Hertz (Hz), indicating the number of periods occurring in a time span of one second. Note that the frequency of a sine wave in Hertz is equal to the inverse of the time span of its period in seconds. A sine wave with a period of 0.5 s will have a frequency of 2 Hz. Note that in the waveform graph, a significant downward spike is observed at a fairly evenly spaced interval. This spacing appears to be the period for the signal in general. The inverse of this interval corresponds to the very first peak on the left in the spectral graph, the fundamental. This relationship indicates that theoretically it is possible to extract fundamental frequency from both the waveform and spectral representations of a sound signal. Note that while the fundamental is the first peak, it is not necessarily the peak with greatest amplitude. Several peaks spaced at roughly equal distances can be seen in the depicted spectrum. These peaks indicate the harmonics.

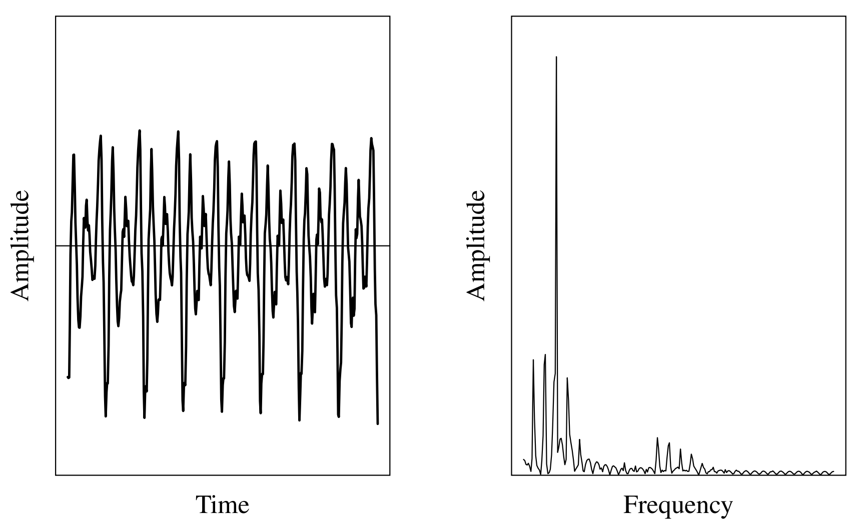


Figure 4-1. Graphs of a periodic signal and the amplitude from the associated spectrum.

Pitch implies a perceptual scaling of the frequency axis. The strings on a piano are tuned to vibrate so that perceived frequency, or pitch, is generally around a targeted frequency specific to each key on the keyboard. In scores, the lines and spaces on a musical staff correspond to keys on the piano. A range of frequencies that perceptually sounds closest to a frequency targeted for tuning can be defined for each key. While the exact ranges for each key are not immediately relevant, it is important to note that they derive from a nonlinear conversion of frequency. For purposes of this discussion, the *fundamental pitch* of a periodic or approximately periodic sound will be defined as the piano key whose specified frequency range encompasses the fundamental frequency appearing in the signal's spectrum. Conversion of fundamental frequency to pitch is essentially a nonlinear quantization of frequency. It is this value that we wish to automatically determine when analyzing pitched portions of signal from a sung performance.

Several methods of detecting fundamental pitch from musical signals have been proposed and investigated. Each of these techniques generally provides a method of analysis to determine relevant spectral content or periodicity in the signal and a decision criterion for determining fundamental pitch. These methods can be categorized into essentially four groups based on the analysis method:

1. Methods based on spectrum computation and examination of harmonics.

2. Methods based on autocorrelation of the signal and degree of match.

3. Methods based on features of the waveform and pattern matching or curve fitting.

4. Methods based on filter banks and examination of amplitude and periodicity.

Complete implementations will often include some form of preprocessing that is intended both to separate pitched portions of the signal from unpitched portions and to emphasize the most relevant or expected frequencies (periodicities). Pitch detectors are also likely to apply methods of both quantizing frequency and smoothing or post-processing of the pitch estimates. In addition to methods for detecting pitch in music, researchers have investigated many methods for detecting fundamental pitch in speech. In general, these detectors also apply one or more analysis methods from the four listed categories.

The first method listed uses a representation of the spectrum. This approach must first calculate an approximation to the spectrum for a region of signal. Typically the spectrum is calculated by using the fast Fourier transform, as presented in Chapter 2. Modified forms of transform are sometimes applied to take advantage of properties that extract only frequency components of interest or to calculate them more expediently (Brown 1991). Commonly the amplitude or power (amplitude squared) is extracted from the Fourier transform and phase information is discarded. However, phase information can also be used to determine frequency (Charpentier 1986), as can the cepstrum (the Fourier transform of the logarithm of the spectrum) of the signal (Noll 1970). Decision criteria for identifying the fundamental frequency generally involve either a direct method for selecting the fundamental (Rabiner *et al*. 1976) or a pattern matching approach that considers the overall harmonic structure (Hermes 1988; Brown 1992; Maher and Beauchamp 1994). These methods generally rely on formal mathematical properties expected of the signal. Recent publications also discuss investigation of more general pattern recognition and statistical techniques for choosing the fundamental (Doval and Rodet 1993; Taylor and Greenhaugh 1993; Taylor and Greenhaugh 1994).

Given an infinite periodic function, if one duplicates the function and shifts the duplicate to the right by one full period, the two functions will again overlap completely. A mathematical result of this property is that the integral of the product of a periodic function and its shifted duplicate will be maximized when the duplicate is shifted by an integer multiple of the period. If one calculates this integral after each of many incrementally larger shifts of the duplicate function, the period of the function would be well-approximated by the size of the shift that maximizes the integral. The method of calculating these values is referred to as autocorrelation:

$$a(n) = \int f(m) \cdot f(m-n)\ \partial m$$

Note that when calculated in closed-form for a continuous function, the autocorrelation provides another continuous function defining the desired value for any real-valued shift size. Methods for expediently approximating the autocorrelation for a sampled function are well-known (Rabiner and Schafer 1978). Since real musical signals are not infinite and periodic, decision criteria used with this method often incorporate additional information from the waveform (Rabiner 1977; Fernandez-Cid and Casajus-Quiros 1994; Casajus-Quiros and Fernandez-Cid 1994).

The third method of detecting fundamental pitch relies on calculating relevant features of the waveform. These features often include information like local minima and maxima in the waveform and the time between amplitude peaks or amplitudes of zero. This information is often mathematically treated to produce an estimate of the pitch (Steiglitz, Winham, and Petzinger 1975; Cooper and Ng 1996). It is also possible to compare this information against templates or other representations of signals having a known pitch. This latter method would be specific to a particular class of musical signals, such as tones produced by a specific instrument or by a specific performer.

The final method of detecting fundamental pitch is based on filtering. Filtering is a method of removing from a signal, or attenuating, frequency components in a specific range. While in concept filtering amounts to multiplying the spectrum by an appropriate function, in reality it is often implemented by direct calculation on the signal. Pitch detection can be based on trying to filter out all harmonics from a signal (Kuhn 1990; Lane 1990). For a truly periodic signal, this produces a sinusoid with period equal to the fundamental frequency. Appropriate decision criteria must be applied to determine the fundamental in real musical signals. Alternatively, measurements of the output of several filters could be considered simultaneously, allowing the use of decision criteria similar to those used on the estimated spectrum or for pattern matching.

Approaches to pitch detection often are tailored for a particular application. Sometimes the analysis methods and decision criteria include information or assumptions specific to a targeted task. For instance, when extracting pitch from an instrument known to have a weak fundamental, the decision criteria may consider the relative amplitude of the second harmonic. Similarly, detecting pitch in recordings may not require real-time processing, allowing for more detailed and complete analysis of the signal than is possible when detecting pitch in live performances. For purposes of this study, the method described by (Kuhn 1990) was applied. This method was selected because it had been applied specifically to vocal performances. The published description carefully addressed problems specific to using the technique to detect pitch in singing. Additionally, the method was easy to implement and was designed for real-time applications. Further, not only had the author used it for an automated transcription system, but it had been used in a vocal accompaniment task as well (Inoue, Hashimoto and Ohteru 1994).

This implementation of Kuhn's approach to pitch detection is based on the use of several lowpass filters. *Lowpass filters* are designed to attenuate frequencies above a certain point, thereby passing only the energy associated with the lower frequencies. The point at which frequency attenuation is intended to start is called the *cutoff frequency*. In actual filter implementations the cutoff frequency specifies where the filter gain is $\frac{1}{\sqrt{2}}$. For the pitch detector, the cutoff frequencies of the filters are spaced by a half-octave. An *octave* is two frequencies related by a factor of two, the higher frequency being twice the lower frequency. A *half-octave* is two frequencies related by a factor of $\sqrt{2}$. Note that since pitch and frequency are logarithmically related, an octave spans twelve piano keys and a half-octave spans six. The purpose of spacing the filters by a half-octave relates to the second harmonic. In a periodic signal, the fundamental and the second harmonic span an octave. Thus, by spacing the cutoff frequencies of lowpass filters at half-octaves across the range of a singer, one or more of the filters will pass energy mainly from the fundamental but not the harmonics. This concept is depicted in Figure 4-2. The output of this filter should be essentially sinusoidal, with a period that is the inverse of the fundamental frequency (recall that a period of 0.5 s equates to a frequency of 2 Hz).

For a pure sinusoid, the distance between any two axis crossings, or *zero-crossings*, equals half the period of that sinusoid. When pitched signal is passed through a filter with a cutoff frequency between the fundamental and the second harmonic, the period (fundamental) can be determined by calculating the average time between every second zero-crossing following the first observed crossing:

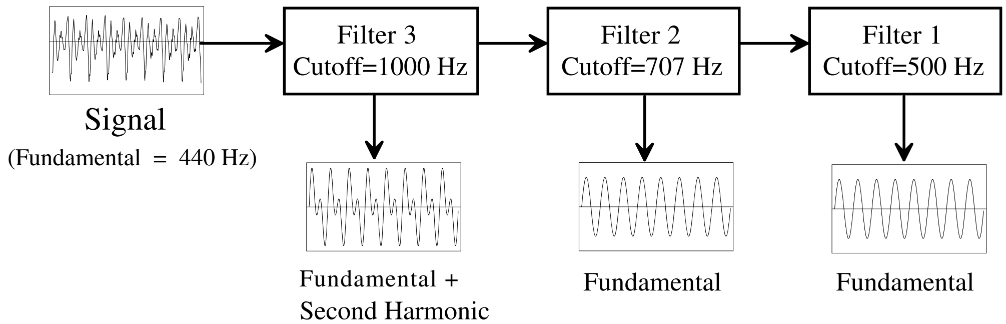


Figure 4-2. A bank of filters whose cutoff frequencies are spaced by half-octave intervals.

$$Period = \frac{TimeOfZero[2n+1] - TimeOfZero[1]}{n},\ n \geq 1$$

The trick to obtaining a reasonable estimate of the fundamental frequency is determining which filter over the range of the singer passes the fundamental but not the harmonics. To do this, Kuhn's approach applies the zero-crossing calculation to the output of each filter and also determines the maximum amplitude of the output of each filter. The filter with lowest cutoff frequency whose output has a maximum amplitude at least 25% of the maximum amplitude of any filter output is then identified. If the frequency estimate (based on zero-crossings) for this filter's output is below the cutoff of the filter, it is taken as the fundamental frequency. Otherwise, if the frequency estimate for the filter with the next highest cutoff frequency is below that filter's cutoff frequency, then that value is taken as the fundamental:

```
For FILTER = 1 TO NUMBER_OF_FILTERS Do
{
    If  (AMPLITUDE[ FILTER ] > MAX_OF_FILTER_AMPLITUDES / 4)
    {
        If  (1.0 / PERIOD[ FILTER ] <= CUTOFF_FREQUENCY[ FILTER ])
            FUNDAMENTAL = 1.0 / PERIOD[ FILTER ];
        Else If  (1.0 / PERIOD[ FILTER+1 ] <= CUTOFF_FREQUENCY [ FILTER+1])
             FUNDAMENTAL = 1.0 / PERIOD[ FILTER+1 ];
         Else
             FUNDAMENTAL = 0.0;
         Exit;
    }
}
```

Since any implemented filter does not completely attenuate frequencies immediately above the cutoff, this decision criteria also considers the filter with the next highest cutoff frequency. This approach can handle cases where significant energy from the fundamental "leaks through" a filter with cutoff frequency at or just below the fundamental. Such attenuated fundamentals can be corrupted by low frequency noise in the signal.

The pitch detector used for this project was implemented in software and designed to run on a personal computer with a sound card. The filter bank implemented for the detector consisted of sixth order lowpass Butterworth filters. These filters have an infinite impulse response (IIR). While Butterworth filters do not have the sharpest possible attenuation near the cutoff frequency, they were chosen because of their smooth gain response. Other filters of the same order but with sharper attenuation often introduce "ripples" in the filter output that might affect zero-crossings. The pitch detector permits the user to specify a desired detection range. Filter cutoff frequencies are spaced by half-octaves starting one half-octave above the center frequency of the lowest pitch in the detection range. The detector applies only enough filters to span the desired detection range. The filters are configured to operate on an input signal sampled using 16 bit PCM (pulse code modulation) at a rate of 16 KHz. This rate is much higher than necessary for pitch detection. Since the highest pitch produced by singers is always less than 1500 Hz, the Nyquist criterion would require a sample rate of only 3 KHz, and good zero-crossing detection could be done with a sample rate of 5 to 8 KHz. However, since the final tracking system will use spectral features related to phonetic content as well as fundamental pitch, the signal is sampled at a rate that retains higher frequencies relevant for vowel identification. To expedite computation for pitch detection, the signal is downsampled by a factor of three once it has been passed through the filter with the highest cutoff frequency.

The sound signal is preprocessed prior to application of the filter bank. This processing is designed to enhance the frequency range of interest and to distinguish pitched signal from unpitched signal. Using an external analog mixer, a bass boost of 15 dB per octave is applied starting just above 220 Hz (the A below middle C) and a low cut of 18 dB per octave is applied starting at 75 Hz. The bass boost is needed to emphasize fundamentals in the lower pitch range of singing which are often weaker than the harmonics. The low cut attenuates any low frequency noise below the pitch range of singers. The sound card is relied upon to execute any required antialias lowpass filtering prior to sampling.

The sampled signal is blocked into buffers at a rate of approximately 30 Hz. Thus for a sample rate of 16 KHz, each buffer contains 533 samples and spans about 33.3 ms of signal. Each buffer is processed in isolation. First, buffers containing pitched signal must be distinguished from buffers containing unpitched signal (consonants) or silence. Each buffer is examined to determine if a significant period of silence is present. A significant period of silence is defined as at least 15 ms of signal with amplitude below a specified threshold. The threshold is a user controlled parameter on the detector and must be adjusted in conjunction with microphone placement and level controls on both the external mixer and sound card. The duration of 15 ms is selected because it is longer than a single pitch period of the lowest note sung by a bass, and it is around or below half the length of the shortest duration consonants ever observed. Thus silence detection must observe at least one full pitch period of low amplitude signal and is sensitive to even short consonants that span two buffers. Buffers for which a significant silence is not detected will likely contain only pitched signal, providing that the threshold and level controls are properly configured. These portions of the signal can be processed for fundamental pitch detection, while the remaining portions can be discarded.

The result of applying the filter bank and decision algorithm is an estimated fundamental frequency. This frequency is converted to pitch by assuming even-tone tuning. This means that the center frequencies for two notes spaced by a semitone (one piano key, black or white) are related by a factor of $\sqrt[12]{2}$. The boundary between two such notes is taken to be related to the center frequency of either note by a factor of $\sqrt[24]{2}$. Semitone boundaries are determined by specifying the center frequency of the A above middle C as 440 Hz, and specifying the A that is *n* registers removed as 440×2^n Hz. Boundaries for other notes in a given register can then be determined relative to the A in that register. A pitch is determined in this way for each processed signal buffer.

Additional post-processing of the detected pitches helps to improve the final estimate of the fundamental pitch. This process determines the median pitch over every three consecutive buffers of processed signal. Median smoothing helps to discard estimates that constitute misrecognition of the second harmonic as the fundamental, result from processing blocks that span pitch changes, or result from the first buffer following unpitched signal before either the actual pitch or the filters have stabilized. It also reduces the variability of estimates due to local pitch deviations that are either accidental or intentional like vibrato. The result of the median smoothing is reported as the most recent detected fundamental pitch. Thus during a sustained tone, the detector reports detected fundamental pitch at a rate of 10 Hz. Note that this rate of reporting observations agrees with the minimum placed on elapsed time in Chapter 3. This minimum elapsed time was important to determining an appropriate sample rate along the dimension of score position, in order to avoid unacceptable numerical errors when approximating the convolution integral.

The entire sequence of signal processing applied by the pitch detector used in this project is depicted in Figure 4-3. It first applies some preprocessing to emphasize the frequencies of interest and to distinguish pitched signal from unpitched signal. The latter step relies on a threshold that must be calibrated by the user. Next, the pitched signal is filtered by a bank of lowpass filters. The maximum amplitude and average pitch period is determined for the output of each filter. Fundamental frequency is estimated by examining this information. Finally, the estimated frequency is converted to pitch and

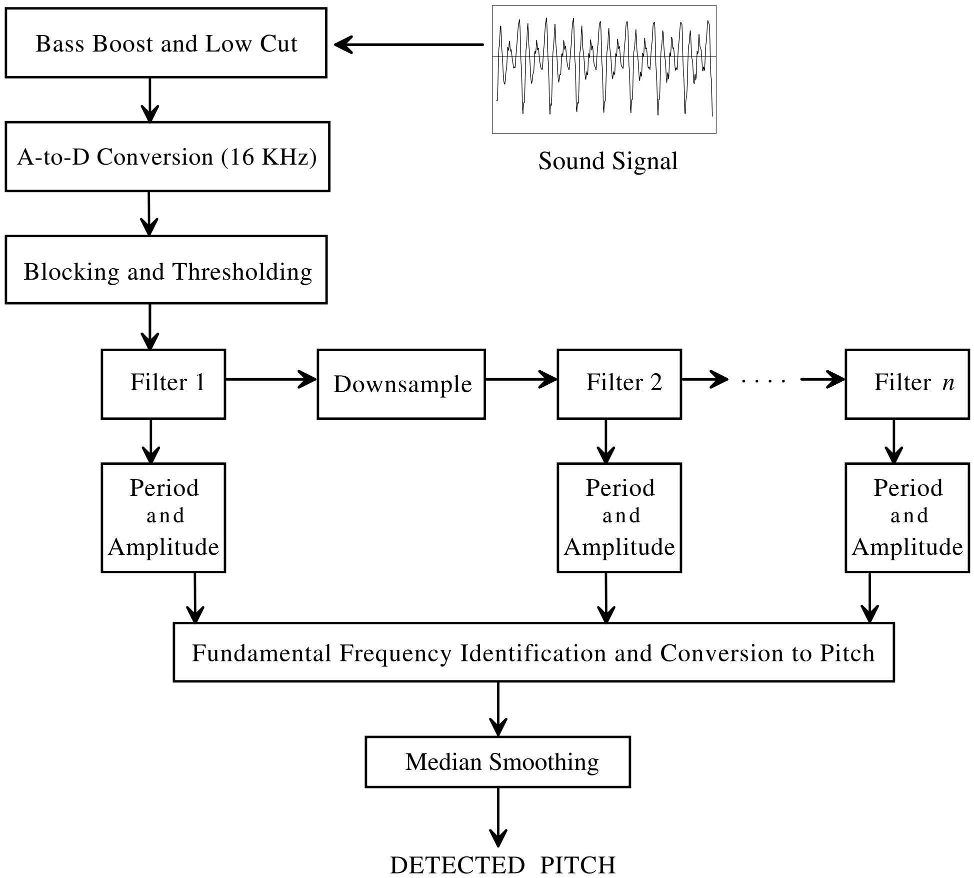


Figure 4-3. Block diagram of the pitch detector used for this project.

median smoothing is applied to remove detection errors and reduce variability of the estimates. For this project, the detector was implemented in software on a personal computer. Running under the Windows NT 4.0™ operating system, processing of a single buffer of signal (~33 ms) requires less than 10 ms total elapsed time on a 66 MHz Pentium processor with buffering of sound card samples handled by the operating system. This detector is used to provide observations of fundamental pitch to the stochastic performance tracking system.

4.4 A Model of Observed Fundamental Pitch

The stochastic score-following model includes a function, $f_{V|I}(v \mid i)$, that specifies the likelihood of making a particular observation given the current score position of the performer. This section describes an approximation to this density function for observations consisting of fundamental pitch automatically extracted from a singer's performance. The approximation is based upon defining each note in the vocalist's part as an event in the score. Additionally, the scored pitch for each event is assumed to be the only significant factor influencing the actual observation. This assumption gives the following approximation:

$$f_{V|I}(v|i) \cong f_{V|ScoredPitch}(v|ScoredPitch(i))$$

While this assumption is not entirely true, scored pitch is certainly an important factor and may be the primary factor in the case of performances by trained singers.

Since the pitch detector is designed to report discrete pitches, a further simplification assumes that the density function is directly proportional to the probability of an observation being $n = v - ScoredPitch(i)$ semitones removed from the scored pitch:

$$f_{V|I}(v|i) \cong f_{V|ScoredPitch}(v|ScoredPitch(i)) \propto \mathrm{P}[v - ScoredPitch(i)]$$

Note that this latter probability distribution is not conditioned on the scored pitch. The distribution is assumed to be sufficiently similar across all pitches so that only a distribution over a single variable, semitone difference, must be modeled. Since the stochastic score-following model normalizes the final score position density function, the probabilities may be directly substituted for density without concern. The assumed constant scaling in the proportionality relationship will be factored out automatically and canceled.

Recorded vocal performances were used to estimate the probability distribution over semitone difference. These recordings consisted of the same eighteen performances used to estimate the density function for actual score distance performed. As previously mentioned in Chapter 3, the recordings included two performances by each of nine singers, spanning all primary voice types both male and female. They included a variety of genres, compositional styles, and performance styles. The recordings were played from a DAT (Digital Audio Tape) into the pitch detection system. The system included the preprocessing components using an external, analog mixer. Tape levels, audio card levels, and detector thresholding were determined by playing each recording through the detector a few times and checking three things. First, the levels were set to avoid clipping of the sound signal. Second, the threshold was set so that all notes triggered the detector (*i.e.*, no soft notes fell below the threshold). Finally, the threshold was set as high as possible to remove many consonants while still triggering on all notes. Each performance was then played into the detector, and the sequence of detected pitches with time stamps was recorded to file.

The sequence of time-stamped pitches for each performance was parsed by determining the first detected pitch after the start of each note. Parsing considered not only changes in the pitch sequence, but more importantly time between reported pitches and, when necessary, graphical displays of the digitized recording and its short-time spectrum. The time between detected pitches is a good indicator of a note boundary. Due to the pitch detector's thresholding, pitches separated by one or more consonants are spaced in time by 150 ms or more. Pitches detected back-to-back during sustained tones will have time stamps separated by less than 120 to 130 ms. This simple criterion can be applied automatically as a first attempt to parse the pitch sequence. Where note boundaries are not clearly indicated by spacing of detected pitches, time-domain and frequency-domain graphs of the digitized recording can be compared against the detected pitches. Pitch changes in the graphs are compared against pitch changes in the sequence, and note durations measured from the graphs are compared against the elapsed time between detected pitches. While not perfect, this approach to parsing the detected pitch sequence is highly accurate given the frequency of reported pitches (one per 100 ms) and the large number of note boundaries easily identified by the time between detector outputs.

Each parsed sequence of detected pitches can be compared against the pitches in the score for that performance. The difference is calculated between the detected pitches spanning each note and the scored pitch for that note, one difference per detected pitch. Distributions for each performance can then be calculated. These per-performance distributions are then averaged to determine an overall distribution for semitone difference. This averaging is done to equalize the contributions from each performance and each singer. The histogram formed in this manner for the semitone differences for all eighteen recorded performances is presented in Figure 4-4. Note that nearly 60% of the detected pitches match identically with the pitch in the score. However, just over 20% of the detected pitches are a semitone below the scored pitch (flat by a semitone), and just over 10% of the detected pitches are a semitone above the scored pitch (sharp by a semitone). Also notice the long and narrow tail on the left side of the graph. This tail continues to narrow but does extend beyond the visible graph. The last bin having any counts appears almost three full octaves flat. A much narrower and virtually invisible tail also extends to the right, but does not continue beyond the visible graph. Histograms for all but one of the individual performances have this same general shape—a prominent peak at zero and a more prominent tail on the left-hand side.

The spread in this distribution can be attributed to several properties of both vocal performance by trained singers and the design of the pitch detector. The vocal techniques of vibrato and expressive performance significantly contribute to the spread in the region around zero semitones (certainly between three semitones flat and three semitones sharp). In extreme cases, expressive performance may contribute to the spread even beyond this range. Vibrato rates of singing are commonly reported in the range of 4 to 8 Hz (Prame 1994; Maher and Beauchamp 1990; Horii 1989), translating to a period of 125 to 250 ms. These time spans are larger than the 100 ms over which median smoothing is applied, implying that the detector output can be affected by vibrato. To verify this situation, the fundamental period of a small

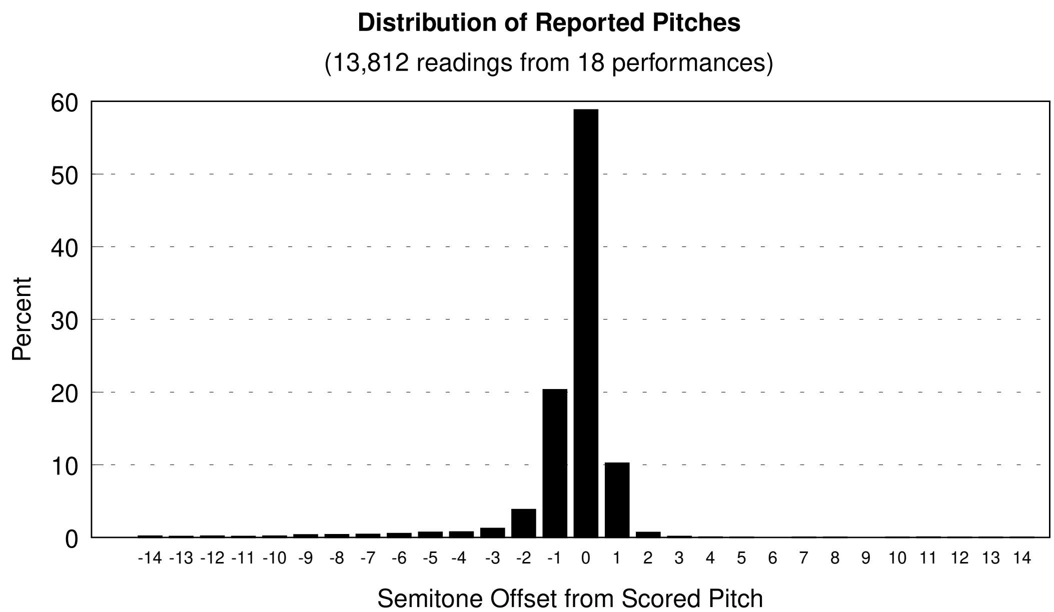


Figure 4-4. Distribution of the difference between detected pitch and the pitch in the musical score.

sample of long duration notes (>500 ms) was measured by examining 100 ms regions of the digitized waveform. Fundamental frequency was converted to pitch by the method described earlier. These few examples confirmed the spread of the distribution near zero. It was noted that the flat readings were often correlated with lower amplitude portions of the signal, while sharp readings were often correlated with higher amplitude regions. This suggests that perhaps the quieter and flat regions of signal must be lengthened in order to compensate for the louder and sharp regions of signal, since human perception of the "central" pitch depends on the energy at certain frequencies over a certain length of time. No confirmation or refutation of this behavior could be found in the relevant literature.

Several properties of the pitch detector may contribute to the distribution's spread and especially to the tails of the distribution. Loud consonants that "leak through" the thresholding contribute to the spread, and so does median smoothing when applied to three consecutive buffers centered on a note boundary. Processing of consonants in particular is likely to produce flat detected pitches. Also, the pitch detector prefers readings from filters with the lowest cutoffs providing the output has significant amplitude. There are likely to be some instances where a weak fundamental is missed and the second harmonic is selected instead. Additionally, if the fundamental is at or just above the cutoff frequency of a filter, it may "leak through" but be attenuated. Such attenuated fundamentals can cause the observed average period of the filter output to be slightly lengthened if low frequency noise is also present in the signal, resulting in a flat detected pitch. Finally, since finite impulse response filters of the type applied do not stabilize instantaneously, the average period of the filter output may be lengthened near the start of a new note, until sufficient steady-state signal has been processed. While median smoothing helps to limit many of these effects, it is likely that these problems still cause some detection errors. Given the types of errors that occur, it is not entirely surprising to see a slightly more prominent tail in the direction of flat pitches versus sharp pitches.

The distribution from Figure 4-4 can be used as an approximation to the observation distribution for fundamental pitch. It provides the probability that the detected pitch will vary from the scored pitch by any number of semitones. The probability values for semitone offsets are presented in Table 4-1. These values will be used in the actual implementation of the vocal tracking system. Note that the table has been simplified in some cases by using only the average likelihood for consecutive offsets having similar likelihood. Semitone offsets beyond a certain range are assigned a very small default probability, indicating that such observations are very unlikely but not assumed to be impossible. This distribution over the difference between detected and scored pitch is used for all events in the score.

As previously mentioned, pitch events correspond one-to-one with notes in the vocalist's part. The length of each event will be determined by the rhythm notated in the score along with nominal tempo markings selected for the different sections of the score. Reasonable nominal tempi can be obtained by selecting a lower bound of all possible tempi for a targeted singer or for any singer who might perform a given score. It is important that the selected tempi be lower bounds in order to avoid numerical errors that can result from undersampling the distance density, as discussed in Chapter 3. Rests in the vocalist's part are associated with a distribution assigning every possible observed pitch an equal probability of 0.0001.

Table 4-1. Probability for $Offset = v - ScoredPitch\ (i)$ approximated from detected pitch sequences for the eighteen recorded performances. Ranges indicate where averaging was applied to offsets with similar likelihood.

Offset	Probability	Offset	Probability	Offset	Probability
< -14	0.0002	-2	0.0383	3	0.0012
-10 to -14	0.0015	-1	0.2033	4	0.0006
-6 to -9	0.0045	0	0.5884	5 to 10	0.0001
-4 to -5	0.0075	1	0.1022	11 to 12	0.0004
-3	0.0125	2	0.0068	> 12	0.0001

4.5 Summary of Observations Based on Fundamental Pitch

The stochastic method for score following incorporates a probability density defining the likelihood of all possible observation values at any score position. This chapter has considered the use of fundamental pitch as one type of observation. The quality of a given observation type can be assessed by considering five criteria: position discrimination, the ease of specifying the distributions (the "regularity" of the distributions), the availability of significant factors affecting the observation, the ease of estimating the distributions, and consistency with the score-following model. Considering these criteria, several problems were noted with using fundamental pitch as an observation type for vocal performances. Methods of pitch detection were then discussed and details of an approach implemented for this project were then presented. This detector was applied to a set of eighteen vocal performances. Using the output of the detector, an approximation to the observation density function, $f_{V|I}(v \mid i)$, for fundamental pitch was determined and discussed in the previous section.

Fundamental pitch in vocal performances is clearly not the ideal type of observation (for numerous reasons). However, it is important not to disregard the discriminatory power of this observation type and the relative regularity of the estimated distribution, even when using a small number of factors. For instance, consider the probability distribution defined by Table 4-1. We can interpret this distribution as both the actual likelihood of observing any detected pitch for a given scored pitch and the likelihood used by the tracking system. The chance that an actual detected pitch for one note in the score will be considered more likely to result from a successor note whose pitch is a semitone lower is thus given by:

$$\sum_{v<ScoredPitch(note)} \mathrm{P}[v - ScoredPitch(note)]$$

The value of this expression is less than 0.30. For the given case, it defines the likelihood of confusing the performed note with a successor based exclusively on a single observation value. This example of a successor one semitone below the performed note is the worst case for a single detected pitch. The likelihood of confusion is smaller for a note with a successor one semitone above (< 0.115), and the likelihood of confusion decreases as the pitch difference between notes increases (< 0.095 for a successor one full tone below). These numbers are potentially large enough to cause errors in position tracking, even when assuming independence of successive observations and when incorporating the distributions for previous position and actual score performed. However, they are not particularly large compared to the same calculations subsequently presented for other observation types.

Much of the statistical modeling presented in this chapter, and in Chapters 2 and 3, focuses on specifying (hopefully reasonable) assumptions that permit for feasible empirical estimation and tractable real-time calculation without unacceptable loss of accuracy. The models presented in this chapter for the observation density and in Chapter 3 for the actual score distance traversed both approximate desired distribution functions by other functions argued to be sufficiently similar. They apply the techniques of both substituting conditioning variables (factors) and directly replacing complex density functions with simpler functions. These "techniques", as well as those applied in Chapter 2, can more generally be viewed as ways of identifying structure or regularity in the true desired functions, permitting them to be well-approximated by less arbitrary functions. These latter functions can be completely specified much more easily in either a closed-form or numerically, using either a completely empirical approach or a combination of empirical and analytical methods.

The approximation approach to obtaining a feasible statistical model facilitates combining multiple sources of information, including symbolic information in the score, real-time measurements of the sound signal, and distributions describing expected or most likely behavior in the real world. This approach has the further advantage that it provides a consistent notation for compactly detailing how the information is combined and the sequence of assumptions or approximations used. The approximations made in Chapter 2 through to the modeling of fundamental pitch are presented in Figure 4-5. This figure completely defines the stochastic score-following model using observed fundamental pitch, estimated rate (tempo), elapsed time, and source position. Only details of implementation (such as the sample rate for score position, the window size, and the method of pitch detection) have been omitted.

As initially mentioned in Chapter 2, there is strong motivation for thinking about a multivariate conditional density function. Specifically, when given a fixed set of both values to be estimated and available, relevant information (the conditioned and conditioning variables), it is not possible to provide better estimates of the unknown values than those given by the conditional distribution. Once the values to estimate and the available information are specified, statistical modeling resolves to obtaining the most accurate approximation to the conditional distribution that is feasible to estimate and tractable to compute. While Figure 4-5 summarizes the approximations used for the score-following model, the question of their accuracy remains. Although the approximation due to uncertainty of previous score position is unavoidable, and the independence assumption is supported through careful design of the sensors, the remaining approximations involve direct substitution of functions and accuracy is a concern. Accuracy assessment in this case requires two distinct questions. First, how accurate are the distributions estimated from collected data; and second, how accurate are the approximations of certain density functions by other functions.

The accuracy of distributions estimated from data is a common concern in classical statistics. This concern is addressed in two ways—through careful design of data collection and sampling, and through calculation of confidence or likelihood of parameter values and functions estimated from the data.

Vocal Performance Tracking Model:

$$f_I^{t1}(i) = f_{I\,|\,J,\,\mathrm{D},\,V}\,(i|\,j = j_0,\, d = d_0,\, v = v_0)$$

$$\cong \int_{j=0}^{\|Score\|} f_{I\,|\,J,\,D,\,V}\,(i|\,j,\, d = d_0,\, v = v_0) \cdot f_{J\,|\,D,\,V}\,(j|\,d = d_0,\, v = v_0)\,\partial j \qquad \text{Uncertainty of } j$$

$$\cong \frac{f_{V\,|\,I}\,(v = v_0|\,i)}{f_V(v = v_0)} \cdot \int_{j=0}^{\|Score\|} f_{I\,|\,J,\,D}\,(i|\,j,\, d = d_0) \cdot f_{J\,|\,D}\,(j|\,d = d_0)\,\partial j \qquad \text{Independence}$$

$$\cong \frac{f_{V\,|\,I}\,(v = v_0|\,i) \cdot C(i,j,d = d_0)}{\int_{i=0}^{\|Score\|} f_{V\,|\,I}\,(v = v_0|\,i) \cdot C(i,j,d = d_0)\,\partial i} \qquad \text{Convolution and Substitution of Source Position}$$

where $C(i,j,d = d_0) = \int_{j=0}^{\|Score\|} f_{I\,|\,J,\,D}\,(i|\,j,\, d = d_0) \cdot f_{J\,|\,D}\,(j|\,d = d_0)\,\partial j$

$$\cong \int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\,d = d_0) \cdot f_I^{t0}(j)\,\partial j$$

j = source position of the performer.
i = destination position of the performer.
d = estimated score distance traversed by the performer.
v = newly reported observation.

Distribution of Actual Score Performed:

$$f_{I-J\,|\,D}(i-j|\,d = d_0) \cong f_{I-J\,|\,R,\,\Delta T}\,(i-j|\,r = r_0,\, \Delta t = \Delta t_0) \cong \frac{1}{(i-j)\sigma\sqrt{2\pi}}\, e^{\frac{-(\ln(i-j)-\mu)^2}{2\sigma^2}}$$

where $\mu = \ln r_0 - \frac{1}{2}\sigma^2 + \ln \Delta t_0$

$$\sigma^2 = \ln\left(\frac{1}{.02948\Delta t_0} + 1\right)$$

$\Delta t_0 = t1 - t0$

r_0 = estimated average rate over preceeding ~3 seconds.

Distribution of Observed Fundamental Pitch:

$$f_{V\,|\,\mathrm{I}}(v = v_0|\,i) \cong f_{V\,|\,\mathrm{ScoredPitch}}\,(v = v_0|\,ScoredPitch(i))$$

$$\propto \mathrm{P}_{SemitoneOffset}[v_0 - ScoredPitch(i)] \cong \text{probabilities from Table 4-1}$$

Figure 4-5. Stochastic score-following model using pitch, rate, elapsed time, and source position.

The former methodology considers how data is to be collected relative to the targeted population (all possible instances) and the desired estimates. It encourages use of techniques that remove or minimize bias and uncertainty, including randomization and reliance on properties of estimation methods based on large sample theory (*e.g.*, the law of large numbers). Data collection done for this project adheres to these policies to the extent possible, and such techniques are argued to have been fairly successful for the estimation done in this chapter and in Chapter 3. The latter methodology attempts to answer how likely or "believable" it is that estimated values or functions are the actual values and functions. It applies analysis and calculations based on specific theoretical assumptions, such as the methods of data collection and possibly properties of the estimated distribution or the true underlying distribution. These techniques were used extensively in Chapter 3 to assess the lognormal convolution model for actual score performed. Due to the simple nature of the estimation done in this chapter and the large number of data points obtained from an arguably unbiased sample, no such techniques have been used to assess the histogram model for observation of fundamental pitch. Note that both methodologies for addressing accuracy of estimation from data contain well-known and fairly comprehensive techniques described in textbooks.

The accuracy of approximating density functions by other functions is a little more challenging to assess. The task is complicated by the fact that the approximations are used to simplify the subsequent estimation and computational processes, so the functions being approximated are not known explicitly. Although complete ignorance of the actual density may preclude calculation of any numerical value characterizing approximation accuracy, it is possible to qualitatively assess accuracy or to compare alternative substitutions through another means. Such assessment can be based upon central limit theorems.

Central limit theorems of various forms exist, but essentially they all state that under various general sets of circumstances, the sum or product of a number of random variables becomes more normal or lognormal as the number of variates increases. The circumstances required for this to be true usually include properties such as no single value in the sum is large enough to dominate the others, the variables are independent or mostly independent, and possibly that the variables are identically or similarly distributed. The convergence to normality occurs more rapidly (requires smaller samples) to the extent that these criteria are satisfied. It is possible to apply central limit theorems to assess the accuracy of density approximation due to the conditioning variables (factors) used in the functions. Specifically, if the primary factors that influence the values of the conditioned variables are characterized by the conditioning variables in the function; then for each assignment of values to the conditioning variables, a central limit theorem will likely apply to the distribution over the conditioned variables. In other words, the assignments of the conditioning variables will essentially determine the value of the conditioned variables except for a sum or product of several smaller and essentially random perturbations. The distribution over the conditioned variables will likely be well-approximated by a normal or lognormal density. Note that this is particularly likely to be true of measurements taken from signals produced in the physical world.

For every fixed assignment of conditioning variables, if the empirically estimated distribution over the conditioned variables is approximately normal or lognormal and has small variance, then it is very likely that the approximation of one function by another is good. Our confidence that this is true is increased to the extent that the significant conditioning variables are known and included in the function, and that the empirical estimation process is conservative (*i.e.*, uses sufficiently large and unbiased samples from the target population). Note that "small variance" can often be interpreted as small relative to the variation of the conditioned variables over the ranges of the conditioning variables. Also, it is important to point out that this criterion for evaluating approximations says nothing about when the approximation is poor. When distributions over conditioned variables clearly are neither normal nor lognormal or have large variance, it is not necessarily safe to conclude that poor approximation is the sole or even primary cause. These situations only indicate that further investigation or estimation is warranted. Finally, the normality criterion can obviously be invalidated by making abusive assumptions when substituting functions. For instance, it is not appropriate to automatically assume normality or lognormality and small variance without assessing these assumptions or providing an applicable and convincing argument based on a central limit theorem.

An assessment of lognormality was undertaken for the distance density modeled in Chapter 3, and a central limit theorem provided. Deviations from lognormality and variances were noted, along with possible estimation errors and relevant conditioning variables not included in the approximation. The estimated distribution shown in Figure 4-4 is not very normal or lognormal. However, it is unimodal and exhibits small variance (the tails drop-off quickly after ± 1 semitone) relative to the distribution of scored pitches. Several factors influencing the distribution were noted earlier, including vibrato and pitch detection applied during or shortly after a consonant in the singing. It is conceivable that explicitly including one or more of these factors in the model might enhance the distribution. For instance, conditioning variables based on detecting the presence of vibrato, and its frequency extent and phase, might provide more normal-looking distributions with smaller variance. Separately modeling erroneous pitches detected during consonants or at consonant to vowel transitions might also help.

This chapter has discussed modeling of a single observation type (specifically fundamental pitch) for use in tracking a vocal performer. The stochastic model for score following based on fundamental pitch is now defined. In the immediately succeeding chapters, two additional observation types are specified and modeled. This modeling extends the score-tracking model to incorporate multiple types of observations. Similar to the discussion of approximation accuracy presented in this chapter, discussion in subsequent chapters addresses how incorporating multiple observation types within the performance tracking model can potentially improve accuracy of position estimation, ultimately leading to a more robust and reliable vocal performance tracking system.

Chapter 5

Events Based on Observing Spectral Envelope

5.1 Extending the Tracking Model for Multiple, Simultaneous Observations

A complete stochastic model for tracking vocal performances was presented in Chapter 4. This model includes estimated distributions for both the actual amount of score performed between observations and the fundamental pitch as reported by pitch detection software at every score position. Both of these distributions were estimated by examining actual vocal performances. The distribution of detected fundamental pitch is substituted for the observation density, $f_{V|I}$, in the general model for stochastic score following presented in Chapter 2. Thus, the completed model of Chapter 4 specifies a performance tracking system based only on observations of fundamental pitch. This chapter describes how to extend the model to incorporate additional observations, and defines a model for tracking vocalists that incorporates observations based on the spectrum of the sound signal in addition to those based on fundamental pitch.

Accurate and reliable identification of a singer's score position requires multiple types of observations, or measurements, of the performance. Chapter 1 provided substantial evidence to support this claim. This requirement motivates extending the score-tracking model to incorporate multiple observation types. As mentioned in the first chapter, it is necessary to observe more than just fundamental pitch. Furthermore, the distribution actually estimated in the previous chapter demonstrated that fundamental pitch is probably not sufficient for determining score position even when also using tempo estimates and elapsed time. Additional observation types based on the phonetic content and amplitude of the sound signal may be helpful (if not necessary) in reliably distinguishing a singer's current score position.

The general model for estimating the score position density can easily be extended to incorporate observations from multiple sensors, providing that two conditions are satisfied. First, the model assumes that observations are independent of the performer's source position and the estimated distance. All observations must satisfy this condition. Second, it must be possible either to specify a joint density for multiple, simultaneous observations or to approximate this joint density by assuming independence. Thus, for two different types of observations, *V1* and *V2*, from different sensors, it must be possible either to directly estimate $f_{V1,V2|I}$ or to use the following approximation:

$$f_{V1,V2 \mid I} = f_{V1 \mid I} \cdot f_{V2 \mid V1, I} \cong f_{V1 \mid I} \cdot f_{V2 \mid I}$$

The joint density can then replace the observation density for a single observation wherever it appears in the general score-following model.

Several considerations influence the decision whether to directly estimate the joint observation density or to approximate it by assuming independence. First, it is important to consider whether or not the value of one observation type is correlated with the value of the other. If so, then direct estimation may provide a more accurate estimate of the true observation distribution than would assuming independence. However, it is also important to consider whether or not the true, primary factors influencing the value of both observation types are known. If so, then substituting these values for conditioning variables (as in Chapter 4) may provide a sufficiently good distribution using the independence assumption. Directly estimating the joint distribution may not offer any improvement, particularly if the relationship between the observation types is purely correlative and not causal. Finally, the time and expense of accurately estimating both forms of the distribution must be considered. For a joint density over several variables, the amount of data necessary to produce a comprehensive and accurate estimate of the density function may be unacceptably large, especially relative to the improvement in accuracy.

Implementation of the tracking model requires consideration of how frequently observations are reported. Since the model represents elapsed time explicitly, it is theoretically possible that model updates may occur at arbitrary time intervals. This property also permits asynchronous reporting of multiple types of observations, allowing revised score position density functions to be generated every time a new observation is available. Even when a joint observation distribution has been directly estimated, any observation types that are not observed at the time of calculation can be removed from the joint distribution through integration. As discussed in previous chapters, omitting observations can affect position estimation. However, because implementing the model requires use of numerical approximations, the permissible elapsed time between calculations of the score position density must meet or exceed a certain minimum. Otherwise, unacceptable errors can result from the repeated, numerical approximation of the convolution integral. Thus in reality, an implementation must guarantee that reports from different sensors are either always synchronized or are separated by at least this minimum elapsed time. This constraint must not be violated when defining multiple observation types for use in a single tracking system. Consequently, all additional observation types defined in this work are designed so that the corresponding observations are generated in synchrony with detection of fundamental pitch.

For purposes of tracking a vocal performer, this project estimated distributions for two types of observations in addition to fundamental pitch—the spectrum of the signal and changes in amplitude. Significant changes in amplitude of a sound signal are often correlated with the beginning of a new note in a performance. Precise definition of these *note onsets* and estimation of an appropriate distribution function is presented in the next chapter. The present chapter deals instead with the spectrum of the sound signal. In particular, since certain characteristics of the spectrum are correlated with the singer's *diction*, or the phonemes that are sung, this chapter describes estimation of distributions over spectral characteristics correlated with the phonetic content of a score.

5.2 Vowel Detection and Measuring Spectral Envelope

The description of fundamental pitch presented in Chapter 4 discussed the logarithmic relationship between pitch and frequency, and presented spectral properties of periodic and quasiperiodic sound signals. These properties included the manifestation of fundamental frequency and harmonics as peaks in the spectrum. Similar to human perception of pitch, the perception of vowels relies upon certain characteristics of the short-time spectrum of the sound. In particular, the location of peaks in certain frequency ranges of the spectrum are important to the recognition of vowels. These peaks are called *formants,* and different vowels often can be distinguished by the frequency of the formants. The spectra for two vowels sung on the same fundamental pitch by the same vocalist are presented in Figure 5-1. Vowels (and all phones in general) are written using special symbols constituting the International Phonetic Alphabet, or IPA. The phonetic spellings of words are often enclosed in brackets. The vowels in Figure 5-1 are similar to the *u* in *push* and the *a* in *father*, and are notated [ʊ] and [ɑ] respectively in the IPA. The narrow peaks corresponding to the fundamental and harmonics are easily seen in each graph. However, the relative prominence of various harmonics is different in each graph. In the graph of the vowel [ʊ], the fundamental is prominent and the harmonic around 1200 Hz is greater than the neighboring harmonics. In the graph of the vowel [ɑ], the third harmonic around 900 Hz is the most prominent and harmonics around 3-4 KHz appear to have more energy than their neighbors. These local maxima in the peaks of the fundamental and harmonics are indicative of the formants.

A precise understanding of formants requires consideration of the human vocal apparatus. Pitched sound is produced when air passes through the *larynx*, causing the *vocal folds* (vocal cords) to oscillate. This oscillation produces quasiperiodic air pulses that travel from the larynx through the *vocal*

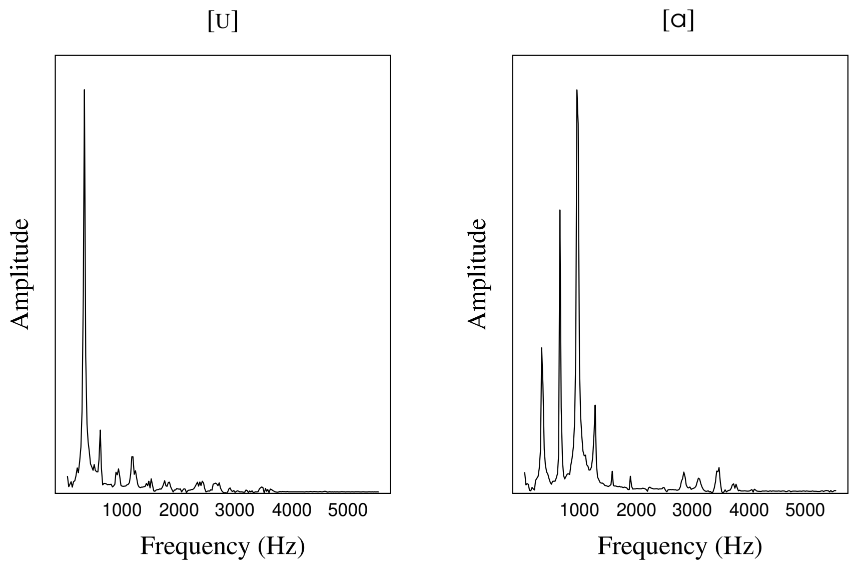


Figure 5-1. Spectra of two vowels, [ʊ] and [ɑ], sung with a fundamental frequency around 300 Hz.

tract, exiting through the mouth or *oral cavity*. The air may also pass through the nasal cavity and exit through the nostrils. In singing, the fundamental and harmonics are produced directly by the vibrations of the vocal folds. The rate of vibration, and hence the change in air pressure, is determined by the tension of the vocal folds. This tension can be controlled directly by muscles in and around the larynx. The quasiperiodic source signal produced by these vibrations has a naturally decaying spectrum, thus higher harmonics are produced with less energy. However, the human vocal tract contains different cavities that act as passive *resonators* to enhance certain frequencies. The frequencies enhanced by these cavities (the *resonance frequencies*) are called *formants*. The size of the cavities, and thus the formant frequencies, change as a singer adjusts different parts of the vocal tract, including the jaw opening, lip opening, and tongue position.

Production of different vowels requires different arrangements of the vocal tract, resulting in different formant frequencies that change the relative strength of the fundamental and harmonics. The vowel quality (or *vowel color*) is determined primarily by the two formants of lowest frequency, and to a lesser extent the third formant. These three formants always lie below 5000 Hz for both male and female singers. Unfortunately, the exact formant frequencies for a given vowel are not uniquely determined. Different singers having vocal tracts of differing length and shape will sing the same vowels with different formants. Formants for two instances of the same vowel sung by the same performer may also differ, especially when the singer changes fundamental pitch and performance style. These changing formants of course complicate modeling of spectral features related to vowels. In addition, even when formants for a given vowel are consistent across instances of that vowel, the changes in fundamental pitch and the harmonics still can change the observed spectrum of the sound signal.

As a first step toward accurate modeling of spectral features correlated with sung vowels, one would ideally like to remove the fundamental and harmonics from the spectrum, retaining only a description of the resonance peaks within the vocal tract. This description would consist of a function characterizing the filtering affects of the vocal tract on the quasiperiodic air pulses produced by the vocal folds. As described in Chapter 4, filtering amounts to emphasizing or attenuating certain frequencies. Thus, filtering corresponds to multiplying the spectrum of a sound signal by a function defining these scalings. This latter function is called the *transfer function*. However, in reality it is not possible to obtain the true vocal tract transfer function from the final sound signal without knowing the precise spectrum generated by the larynx and the transfer function from the singer to the microphone. In general, this information cannot be formulated precisely.

As an approximation to the transfer function, we can use instead a representation of the *spectral envelope*, or a smooth curve connecting the fundamental and harmonics. Several approaches to generating such a representation are possible. A simple approach would be direct sampling of the spectrum, attempting to retain the peaks. Possibilities include retention of local maxima, retention of the peak amplitudes or energy in fixed frequency regions (*e.g.*, apply a filter bank)(Dautrich, Rabiner, and Martin 1983), and simply downsampling the spectrum. A more sophisticated approach involves linear predictive coding (LPC) analysis of the signal (Atal and Hanauer 1971; Markel and Gray 1976). The LPC model assumes that a given speech sample is equal to a linear combination of some fixed number of preceding samples plus an excitation value, essentially an error term. Coefficients for the linear combination can be generated by minimizing the mean squared prediction error when fitting the model to several samples within a small portion of signal. These coefficients define a filter applied in the time domain (*i.e.*, one function participating in a convolution), and the Fourier transform of these coefficients therefore provides an approximation of the vocal tract transfer function (or an approximation to the spectral envelope).

While these representations of the spectral envelope can provide approximations of the vocal tract transfer function, the sound signal must undergo significant additional processing. This processing includes several steps that reduce variability of signals associated with the same vowel. It also alters the representation of the spectrum in ways similar to the processing done by the human ear. First, the signal is often highpass filtered. Since the spectrum of singing naturally decays with increasing frequency, highpass filtering helps to level the spectral envelope. When comparing spectra or applying metrics across a spectrum, this filtering helps to equalize contributions from measurements taken at different points along the frequency range. Second, the logarithm of the spectral envelope is often used instead of the spectral envelope itself (Gray and Markel 1976). Perceived, relative changes in loudness are known to correspond approximately to changes of similar proportion in the logarithm of the amplitude (Benade 1976). Thus, relative comparisons and difference measures applied to the logarithm of the spectrum more closely approximate human perception of changes in loudness. Third, normalization of the spectrum and smoothing along both the frequency and time dimensions can be applied. These techniques reduce variability that results from both changes in loudness and short-term, random spectral fluctuations that are not caused by change of phonation.

Finally, human perception of changes in both loudness and frequency is not constant across all frequencies. For instance, the minimum perceivable change in frequency increases as the base frequency increases. In fact, above 1000 Hz this relationship is approximately logarithmic. Thus, representations of the spectral envelope are often transformed from a linear frequency scale to a logarithmic scale. A special instance of such a transformation based upon perceptual studies is the *mel scale* (Stevens and Volkmann 1940). Similarly, studies have indicated that both perceived changes in loudness and speech intelligibility are affected by bandwidth (*i.e.*, the frequency range) of the signal, according to *critical bands* within the frequency range of human hearing (Zwicker, Flottorp, Stevens 1957; French and Steinberg 1947). A sound signal may be perceived quite differently with respect to loudness or vowel quality when its bandwidth is altered to span additional critical bands or fall within different critical bands. Sampling of the spectral envelope according to such bands and, alternatively, filtering over these bands are two approaches to providing a spectral representation that more closely approximates human perception.

In addition to the LPC coefficients previously mentioned, one set of related coefficients popular for speech recognition is the cepstrum (Furui 1986). The cepstrum is the Fourier transform of the logarithm of the magnitude spectrum. The low order cepstral coefficients define a smoothed spectrum that approximates the vocal tract transfer function, similar to the LPC coefficients. The cepstrum can be calculated from the LPC coefficients, the log spectrum, or the log spectrum converted to a transformed frequency scale like the mel scale. Davis and Mermelstein (1980) provide a comparison of using different spectral representations for speech recognition, and demonstrate improved performance using the cepstrum. Many recent speech recognition systems use the cepstrum (Lee, *et al.* 1989; Austin, *et al.* 1989; Lee 1990). In addition, the vocal accompaniment system constructed by Inoue and colleagues (1994) applied cepstral *liftering* (multiplying the cepstrum by a cepstral "transfer function" or *lifter*) to generate a smoothed spectral envelope.

Representations of the spectral envelope often are discussed in relation to automated speech processing or recognition, and less frequently in relation to reproduction or modeling of singing. While speech and singing are similar, it is important to recognize that several significant differences do exist, particular when the singing is in the operatic or Western classical style. First, fundamental pitch spans a larger range in singing than in average American speech. The trained soprano can produce a fundamental at or beyond 1000 Hz, and the upper half of her range typically begins around 500 Hz. In contrast, speech produced by American females is commonly reported to contain an average fundamental around 200 Hz, and rarely contains a fundamental exceeding 300 Hz (Hollien, Hollien and DeJong 1997). The average fundamental produced by male speakers is reported around 120 Hz, while the range of a trained tenor typically exceeds 400 Hz. Second, the duration of vowels in singing is typically longer than vowels in speech. Spoken vowels are reported to have average durations around 200 to 300 ms and seldom exceed 500 ms (Deng, Lenning and Mermelstein 1989), while note durations often exceed 500 ms and may span several seconds. Third, formant positions often differ between spoken and sung vowels, including the position of the first and second formants that are most responsible for perceived vowel quality. These changes can be intentional as well as accidental, or simply a secondary effect of singing style. For instance, professional sopranos are known to adjust the location of their formants as the fundamental pitch changes (Sundberg 1975) and depending on whether they are singing as a soloist or within a choir (Rossing, Sundberg, and Ternström 1987). It is well known and accepted that these changes may cause alteration of the vowel quality, even to the point where the vowel identity is lost (Vennard 1967). Male operatic singers exhibit what is called a "singer's formant", where the fourth and fifth formants are drawn much closer to the third formant than they are in normal speech (Bartholomew 1934; Sundberg 1974). This alteration of the spectrum is known to be significant to the perceptual distinction of operatic singing from speech (Sundberg 1987).

The representation of the spectral envelope used for this project is based upon a straightforward combination of many of the previously discussed transformations. For simplicity, this method is applied to the same sound signal used for pitch detection. This signal is sampled using 16 bit PCM (pulse code modulation) at a rate of 16 KHz. This sampling rate is sufficient to capture frequencies at or below 5 KHz, a range spanning the formants that determine vowel quality. Sampling is done by a PC sound card with appropriate antialias filtering. Recall also that a bass boost (starting around 220 Hz) and low cut (starting at 75 Hz) are applied to this signal by an external mixer. The signal is blocked into buffers at a rate around 30 Hz, each buffer spanning approximately 33 ms. Buffers are then identified as containing pitched or unpitched signal according to the thresholding mechanism described in Chapter 4. This thresholding is retained for spectral envelope estimation in order to align observations of spectral envelope with observations of fundamental pitch. Consequently, spectral envelope observations are made only once for every 3 signal buffers, or at a rate of 10 Hz. Since our interest is in modeling spectral envelope primarily for sung vowels, the use of the threshold and an observation rate of 10 Hz is not problematic.

As mentioned in Chapter 4, a fundamental pitch is determined for all blocks identified as containing pitched signal. The median over every 3 consecutive blocks is reported as the observed fundamental pitch. To efficiently estimate the spectral envelope, only every third block is processed (*i.e.*, the most recent block at the time a median pitch is reported). The signal buffer for the block is duplicated and processed independently for spectral envelope and fundamental pitch. To obtain the envelope, the following first order highpass filter is applied to the signal initially:

$$s[n] - 0.95\, s[n-1]$$

where s represents the signal. Next, the filtered signal is divided into four distinct regions spanning around 8 ms each. The following Hamming window is applied to each region:

$$s[r_i + n] \times \left(0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)\right),\ 0 \le n \le N-1$$

where N equals the number of samples in each of the four regions and r_i indicates the start of each region. This windowing gently tapers the signal regions to zero at both boundaries, reducing problems that may result when computing the Fourier transform at discontinuities. Next, as expected, the spectrum of each signal region is computed using the fast Fourier transform. Since the signal is sampled at 16 KHz, the frequencies in the calculated spectrum will range from -8 to 8 KHz. Each signal region is padded with zeros in order to obtain a 512 point transform, providing spectrum samples spaced by $\frac{16000}{512} = 31.25$ Hz.

The samples between 200 and 5300 Hz are extracted from each computed spectrum. This frequency range encompasses the most relevant frequencies for assessing vowel quality. For each sample in this range, the logarithm of the magnitude squared is computed. This transformation provides a logarithmic conversion of a loudness measure (power) at each sampled frequency. These log power measures are then normalized across all samples of interest, for each of the four regions. The log power spectra across all four regions are then averaged. Computing an average over multiple spectra for small regions of signal produces a smoother log spectrum than computing a single spectrum over a longer signal region. Figure 5-2 demonstrates this property by showing two versions of the spectrum computed from the same signal. The smoother spectrum is preferred for characterizing the vocal tract transfer function independent of the fundamental and harmonics.

The frequency range from 200 to 5300 Hz is partitioned into intervals spanning roughly one-third of an octave. The partition boundaries are computed by taking the first sample below 200 Hz (187.5 Hz) and successively multiplying by 1.25, which is just less than $\sqrt[3]{2}$. These boundaries define a total of 15 partitions over the frequency range of interest. Finally, the maximum log power is selected from all samples within each partition of the averaged spectrum, producing a final representation of the spectral envelope that contains 15 log power readings. This representation is of lower dimension than the

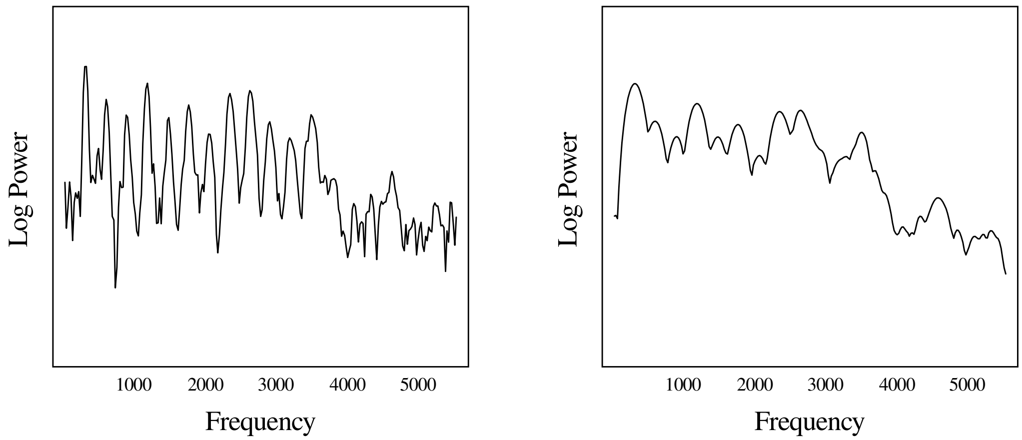


Figure 5-2. Two log spectra calculated from the same 33 ms of signal from the vowel [U] sung with a fundamental around 300 Hz. The left graph depicts a single 512 point FFT calculated over the entire signal; the right graph depicts the average of four 512 point FFT's calculated over 8 ms regions of the signal. A highpass filter and a Hamming window were applied prior to each FFT. Note the increased resolution of the fundamental and harmonics in the graph on the left.

computed discrete spectrum, and essentially captures the most prominent harmonic (when harmonics are present) within each of fifteen equal length regions on a logarithmic frequency scale. Figure 5-3 shows three graphs depicting various representations of the spectrum, including the maximum log power over the 15 partitions, for an example of a sung vowel [U]. Note that the final representation retains an indication of the formants while removing most spectral effects specific to the fundamental and harmonics.

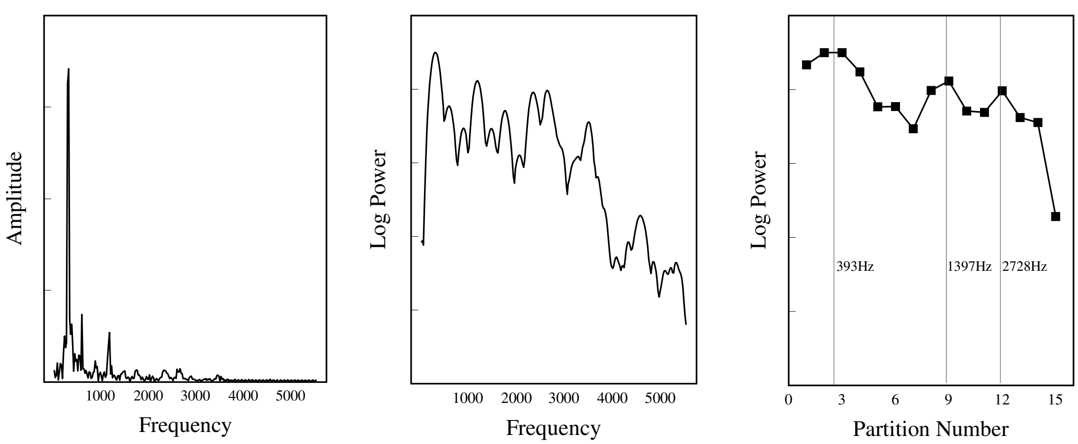


Figure 5-3. Three graphs showing different versions of the spectrum for the same portion of signal from the vowel [U] sung with a fundamental around 300 Hz. From left to right, 1) the magnitude spectrum over the frequency range of interest for 33 ms of signal, 2) the average log power spectrum after applying a highpass filter and a Hamming window to four 8 ms regions of signal, and 3) the highest value from the second graph in each of 15 partitions of the frequency axis. Each partition spans approximately one-third of an octave. The frequency values on the third graph indicate the upper boundary of the partitions spanning each of the first three formants.

The entire sequence of signal processing for extraction of the spectral envelope is depicted in Figure 5-4. The same signal is shared for both pitch detection and spectral envelope extraction through to the blocking and thresholding described in Chapter 4. At this point, only the third of every three consecutive blocks that pass the threshold is processed for spectral envelope. The signal block is divided into four distinct regions each spanning about 8 ms. Each region is highpass filtered and Hamming

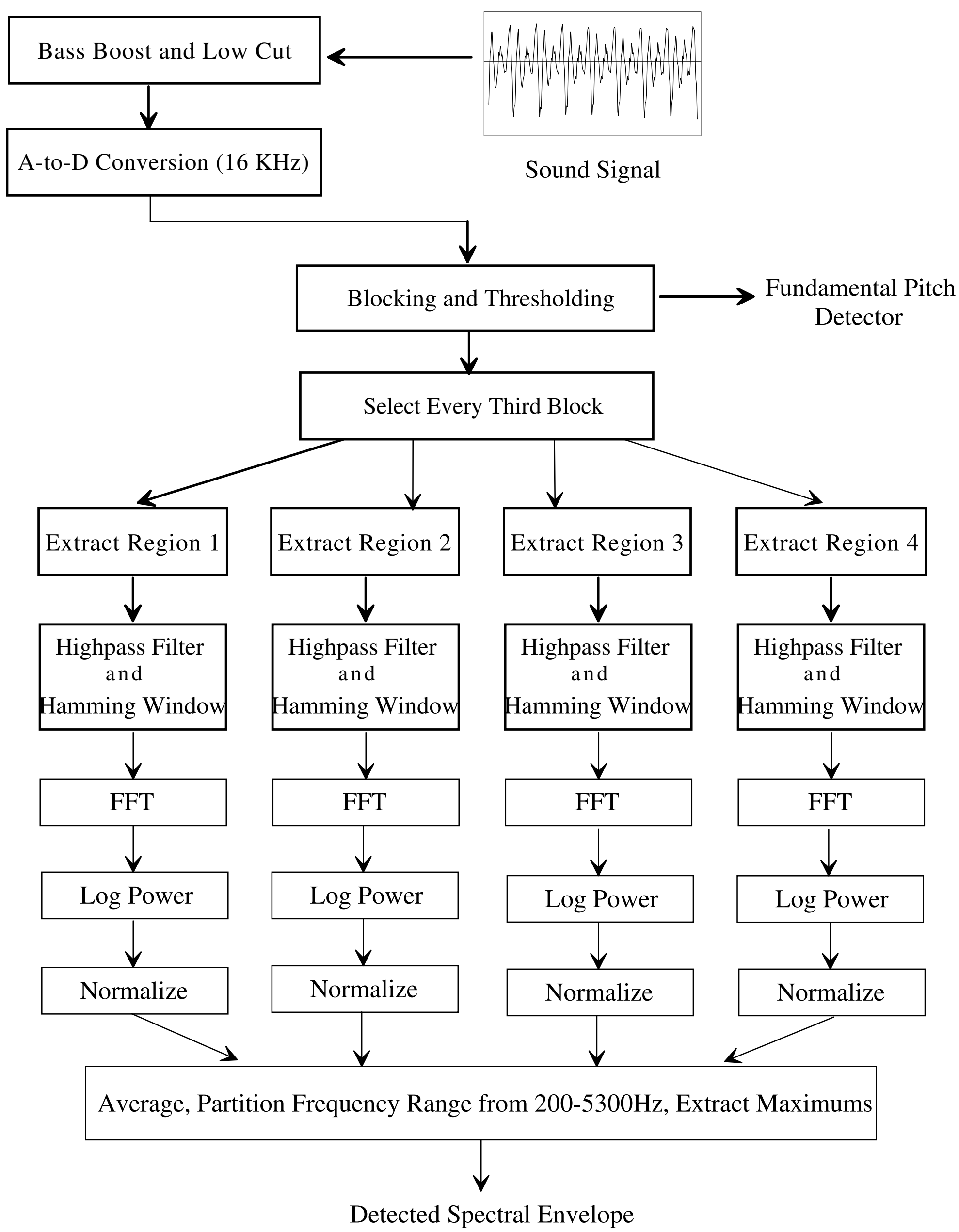


Figure 5-4. Block diagram of the spectral envelope detector used for this project.

windowed, and a normalized log power spectrum is computed. The four spectra are then averaged and the frequency range of interest is partitioned into regions spanning one-third of an octave. The maximum log power from each partition is extracted to produce the final representation of the spectral envelope. Like the pitch detector, the spectral envelope detector was implemented in software running under the Windows NT 4.0™ operating system. Processing of a single block of signal (~33 ms) extracting both fundamental pitch and spectral envelope requires less than 30 ms total elapsed time on a 66 MHz Pentium processor, with buffering of sound card samples handled by the operating system. This detector is used to provide observations of spectral envelope to the stochastic performance tracking system.

5.3 Spectral Envelope in Vocal Performances

In Chapter 4, five criteria that an ideal observation type should satisfy were presented. These five criteria consist of the following: position discrimination, the ease of specifying the distributions, the availability of significant factors affecting the observation, the ease of estimating the distributions, and consistency with the score-following model. While criteria for an ideal type of observation have been provided, realistic types of observations inevitably fail to satisfy these requirements completely. Similar to fundamental pitch in vocal performances, the selected representation of the spectral envelope is no exception. The definition of this representation provided in the previous section makes clear several shortcomings of using spectral envelope.

First, the spectral envelope cannot completely discriminate score position. Recall that data presented in Chapter 1 showed that in vocal music, a change of syllable does not occur on every note. Several successive notes may be sung on the same vowel. It is not uncommon to find vocal scores that require the vocalist to sing a single vowel across several measures, sometimes containing fifty or more notes. To the extent that the chosen representation of the spectral envelope is consistent during a single vowel, it will not assist in discriminating score position within these regions. In addition, this representation provides a low resolution view of the spectrum that may not always distinguish different vowels with similar formants sung with the same fundamental pitch. Depending upon the degree of similarity between the vocal tract transfer functions for these vowels, humans also may have difficulty distinguishing them.

As an observation type, spectral envelope does not permit easy specification of distributions for every point in the score. Several factors hinder precise definition of the distributions. First, unlike changes in pitch which are precisely and explicitly notated in musical scores, the transitions between phonemes are not clearly indicated. Although syllables are explicitly associated with notes, no durations (either relative or absolute) are indicated for individual phonemes. Musical scores provide no guidance regarding the positioning of *diphthongs* (a smooth transition from one vowel to another vowel on the same syllable) or *voiced consonants* (consonants that are pitched, such as [l] as in *lute* and [m] as in *mute*) sung on the same note. Whether leading consonants in the syllable should be sung preceding the note or as part of the note is unspecified as well. Second, although the thresholding applied by the spectral envelope detector is intended to distinguish pitched and unpitched signal, it is not always successful. Consequently, the reported spectral envelope sometimes is derived from loud *unvoiced consonants* (consonants that are unpitched, such as [t] as in *toot* and [f] as in *flute*). It is necessary to define distributions for spectral envelope of these consonants as well as for vowels and voiced consonants. The duration and positioning of these consonants is also problematic. Finally, although the discussion of spectral envelope in the previous section implicitly assumed that the vocal tract transfer function remains constant during production of a single vowel, in reality this assumption is not valid. Transitions from one phoneme to another are not instantaneous, so the spectral envelope at the beginning and end of a vowel may differ from the envelope during the middle of the vowel.

Observations of spectral envelope depend on a significant number of factors. First, the observed spectral envelope certainly depends upon the exact phonetic content of the performance. Although operatic singers are trained in proper pronunciation of individual phonemes or *diction*, the phonetic transcription for the lyrics in a musical score can vary from singer to singer. While trained singers are capable of producing the phonetic transcriptions, it may not be reasonable to require them to provide an accurate transcription for every piece. Second, it is well known that the context of the phoneme can affect pronunciation. The tongue, jaw, and lip positions will change depending upon the phonemes that precede and follow a given instance of a phoneme. Since operatic singers perform pieces in a variety of languages, there can be numerous contexts for multiple instances of a single phoneme that appears in multiple languages. Third, as previously mentioned, the spectral envelope may differ significantly for the same phoneme sung with different fundamental pitches. Figure 5-5 shows log spectra for the vowel [u] (the *u* as in *flute*) sung with different fundamentals within the same performance. Substantial differences in the spectra due to the changing fundamental are clearly evident. Finally, spectral envelopes for the same word or phoneme can differ due to musical style, emphasis, and performance technique. The list of possible alterations to spectral envelope is extensive, and no enumeration of all the factors influencing these changes is readily available. Even if such knowledge could be obtained, it is unlikely that all important factors could be automatically measured or made directly available to a computer tracking system. These considerations make it unlikely that all important factors will be available for observations based on spectral envelope.

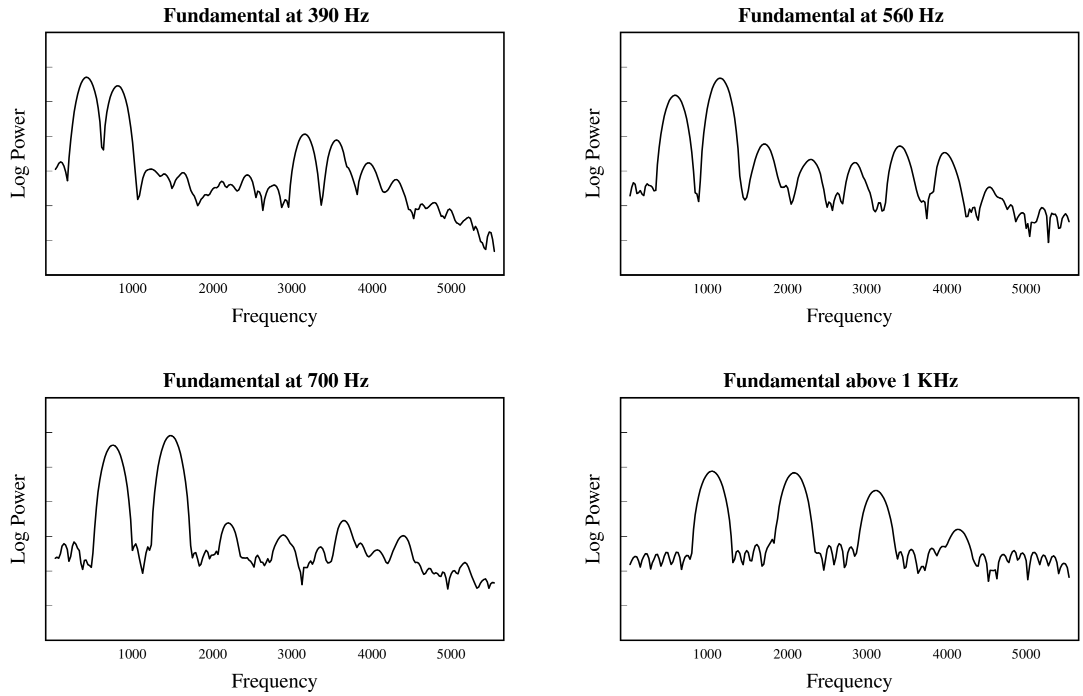


Figure 5-5. Average log power spectra for the vowel [u] sung with four different fundamentals. All spectra were obtained from the same performance. Note the changes in the spectra as the fundamental increases. When the fundamental exceeds 1 KHz, the spectrum consists of only prominent and sparse wideband harmonics.

The large number of factors influencing the observed spectral envelope hinders data collection for this observation type. Lack of sufficient data to model spectral envelope accurately based on phoneme and phonetic context is a widely recognized problem in speech recognition. Expanding the important factors to include fundamental pitch, as well as difficult to define influences like musical style and technique, would require an even larger number of examples. In addition, the difficulties in collecting examples of operatic singing (as discussed in Chapter 4) further hinder data collection for modeling observed spectral envelope. Consequently, it is unlikely that a model including all the previously identified factors could be estimated adequately.

Finally, observations of spectral envelope may not entirely satisfy the assumptions of the score-following model. Although the chosen representation of spectral envelope can be reported as a fixed value (vector), it is not valid to assume that these observations are entirely independent of the previous position of the performer and the estimated amount of score performed. Since the amount of signal processed will span a fixed duration, changes in a singer's tempo can alter the portion of the performance that is processed when detecting spectral envelope. Thus the observation distribution may differ depending on the actual tempo, especially for score positions near phoneme transitions. Actual tempo significantly influences both estimated tempo and previous position. However, this problem may not be as significant for observations of spectral envelope as compared to observations of fundamental pitch, primarily because the amount of signal used to estimate the envelope (about 33 ms) is much less than the amount of signal used to estimate pitch (about 100 ms). Consequently, the variance of the envelope observations for a given position may not be influenced as significantly by changes in tempo.

Just like observations based on fundamental pitch, observations based on spectral envelope fail to completely satisfy the criteria for an ideal type of observation. While the two types of observations satisfy the various criteria to different degrees, neither appears sufficient to guarantee successful vocal performance tracking if used alone. In reality, however, adequate and robust tracking might be accomplished using reasonable approximations and adequate estimates of the observation distributions for either type of observation. Support for such approximations requires empirical investigation. In addition, it has been posited that simultaneously observing both fundamental pitch and spectral envelope, and combining the respective distributions as discussed in the first section of this chapter, can yield even better tracking. This claim is examined through empirical investigation detailed in subsequent chapters.

5.4 A Model of Spectral Envelope

The stochastic score-following model includes a function, $f_{V|I}(v \mid i)$, that specifies the likelihood of making a particular observation given the current score position of the performer. This section describes an approximation to this density function for observations consisting of estimated spectral envelope automatically extracted from a singer's performance. Ideally, this approximation should be developed by including all primary factors influencing the observed spectral envelope. Then, when combining observations of fundamental pitch and spectral envelope in a single model, it would be valid to assume that either the two types of observations are probabilistically independent or the product of the estimated distribution functions accurately approximates the joint distribution:

$$f_{Pitch,Envelope|I} = f_{Pitch|ScoredPitch(I)} \cdot f_{Envelope|EnvelopeFactors[1\ldots n]}$$

Note that the primary factors determining the spectral envelope may or may not include the observed fundamental pitch.

One important limitation on the number of factors actually considered is the amount of available data for estimating the distribution. Recorded vocal performances were used to estimate the probability distribution for spectral envelope. These recordings consisted of the same eighteen performances used to estimate the density function for observed fundamental pitch. As previously mentioned, the recordings included two performances by each of nine singers, spanning all primary voice types both male and female. They included a variety of genres, compositional styles, and performance styles. The pieces contained lyrics in either English, Italian, or German. Although this set of recordings provided sufficient examples for estimating the distribution of fundamental pitch based simply on the semitone difference between observed and scored pitch, it did not provide sufficient data to estimate spectral envelope even based exclusively on the expected phoneme. In particular, certain phonemes did not appear with sufficient frequency. Adding additional factors to the model, such as fundamental pitch or phonetic context, would only worsen the data deficiency. Estimates from the available data indicated that at least twice the number of recorded performances were needed to rectify this situation. Based on the time needed to collect the recordings actually used in this study, doubling the number of recordings would require several additional months of data collection if not more. Thus, an initial approximation to the spectral envelope density used phonetic information as the only factor. Several restrictions were applied to the modeling of this data, in an attempt to reduce estimation error.

The distribution of observed spectral envelope is therefore based upon events that correspond to phonemes in the score:

$$f_{V|I}(v|i) \cong f_{Envelope|ScoredPhoneme}(v|ScoredPhoneme(i))$$

To define this distribution, it is necessary to specify the possible phonemes. The recorded performances contain lyrics in either English, Italian, or German. Distributions were developed for all phonemes appearing in these three languages. Fortunately, singers commonly receive training in diction for all three languages, so the different vowels for the languages are well documented. Texts on singer's diction for English (Marshall 1953; Uris 1971), Italian (Moriarty 1975; Colorni 1996), and German (Cox 1970; Odom 1981) are available. Although texts for the same language do contain differences in both the specified phonemes and the details of pronunciation, these discrepancies are minimal and should not be expected to affect the statistical modeling.

Since vowels are of primary importance to the modeling, care was taken to account for many important variations in vowel quality. Table 5-1 presents the comprehensive list of vowels and diphthongs specified for each of the languages. Several aspects of this list are worth noting. First, only one distribution was developed per IPA symbol, including symbols that appear for multiple languages. The diction of trained singers for a given phoneme was not expected to differ significantly due to language. The limited amount of available performance data further encouraged this modeling decision. Second, the IPA symbol [ɑ] was used to represent both the vowels [a] and [ɑ] in English. The two vowels are distinguished by location of the tongue, which is low and front for [a] and low and back for [ɑ]. The appearance of [a] versus [ɑ] often depends upon regional dialect in spoken American English. Singers are generally encouraged to use [ɑ] when singing English, just as in Italian and German. In

Table 5-1. Vowels and diphthongs that were modeled for each language.

Language	Modeled Vowels and Diphthongs	Number
English	[i][ɪ][eɪ][ɛ][æ][ɑ][ʌ][ə][ɜ][ɔ][ʊ][u][oʊ][au][ʌu][aɪ][ʌɪ][ɔɪ][ju]	19
Italian	[i][e][ɛ][ɑ][ɔ][o][u]	7
German	[i][y][ɪ][ʏ][e][ø][ɛ][œ][ə][ɑ][ɔ][o][ʊ][u][ɑe][ɑo][ɔø]	17

spoken German, the usage is said to depend upon whether the vowel is short or long. This distinction generally does not appear in singing. In Italian, the vowel [a] is not present. Thus for this project, no distinction is made between [a] and [ɑ] for the modeling of spectral envelope.

Finally, diphthongs for both English and German were modeled explicitly, but no diphthongs were modeled for Italian. In Italian, all diphthongs are *textual*, whereby the two component vowels correspond to explicit characters in the text. There is not universal agreement as to whether all instances of two successive vowels in Italian form a diphthong. Consequently, composers throughout the centuries have sometimes scored textual diphthongs as a single syllable for a single note, and sometimes scored them as two distinct vowels on different notes. For instance, the word *mio* is sometimes scored for a single note and sometimes scored as *mi-o* over two notes. To simplify modeling, distributions were estimated only for monophthongs in Italian. In cases where a diphthong is scored for a single note, both vowels are represented explicitly. The vowel most appropriately shortened for performance is treated as a consonant when determining relative duration. Details of calculating durations will be discussed subsequently.

Table 5-2 presents the comprehensive list of consonants specified across all three languages. As with vowels, a single distribution was estimated for each unique IPA symbol. Recall that the envelope detector tries to avoid processing consonants by applying a threshold to the sound signal. This thresholding implies that fewer observations should be expected during production of consonants. Therefore they are less important for observations than vowels, and consequently less care was taken to distinguish consonant quality for distribution estimation. For instance, the dental [d] in Italian versus the alveolar [d] in English and German are not distinguished. In addition, some consonants actually present in the languages have been omitted. These consonants are expected to appear infrequently, to produce observations infrequently, or to be sufficiently similar to other consonants so that a shared distribution representing both phonemes is not problematic. Examples include the German [pf] and the flapped [r].

Table 5-2. Consonants that were modeled for each language.

Language	Modeled Consonants	Number
English	[b][tʃ][d][f][g][h][dʒ][k][l][l̩][m][n][ŋ][p][r][s][ʃ][t][θ][ð][v][w][j][z][ʒ][ʔ]	26
Italian	[b][d][dz][dʒ][f][g][j][k][l][m][n][ŋ][p][r][s][ʃ][t][tʃ][ts][v][w][z]	22
German	[b][ts][tʃ][d][f][g][h][ç][χ][k][l][m][n][ŋ][p][r][s][ʃ][t][v][j][z][ʒ][ʔ]	24

The representation of spectral envelope developed in the previous section consists of fifteen samples from the log power spectrum. Estimating an arbitrary joint distribution over fifteen variables requires a large number of data points and can require a representation that is difficult to implement. Two approaches to dealing with this problem have been applied for automated speech recognition. One is to fit continuous densities or mixtures of continuous densities, making assumptions about the shape of the distribution or the components in the mixture (Bahl *et al.* 1987; Huang and Jack 1989). Another approach is to develop a multidimensional quantization of the observed vectors and then estimate distributions over the quantized values (Rabiner, Levinson, and Sondhi 1983). This approach is referred to as *vector quantization*, and the quantized values are referred to as a *codebook*. Distributions fit to a manageable number of quantized values require fewer data points for accurate estimation. However, quantization of course introduces an additional source of approximation error. Careful vector quantization for distribution estimation therefore requires considering the tradeoff between estimation error due to quantization and estimation error due to sparse data, analogous to determining an appropriate bin size for a histogram.

Well-known methods exist for generating vector codebooks of fixed size to minimize quantization error on a sample of vectors. The basic algorithm (the generalized Lloyd algorithm) selects a random set of vectors for initial codebook entries, assigns each vector in the sample to one of the codebook entries according to some distance measure, and replaces each codebook entry by the centroid of the vectors assigned to that entry (Linde, Buzo and Gray 1980). This process is repeated multiple times, generally until the average distance between each vector in the sample and its assigned codebook entry falls below a preset threshold. Gray (1984) provides a detailed overview of vector quantization and variations on the basic codebook generation algorithm. The procedure that was used to generate a vector codebook for the spectral envelope vectors is the binary split algorithm. It produces a series of codebooks using the generalized Lloyd algorithm, with each successive codebook having twice the number of entries. The initial codebook consists of the centroid of all vectors in the sample. New codebooks are generated by splitting each entry in the smaller codebook. To split each entry, a very small increment is added to the elements in the vector. The generalized Lloyd algorithm is then applied to the new codebook to minimize the quantization error. For the sample of spectral envelope vectors, a simple Euclidean distance was used to measure the distance between codebook entries and the vectors.

Recorded vocal performances were used to generate the vector codebook entries and to estimate the probability distributions for each phoneme. These recordings consisted of the same eighteen performances used to estimate the distribution for fundamental pitch. The recordings were played from a DAT (Digital Audio Tape) into the spectral envelope detection system. Tape levels, audio card levels, and detector thresholding were set to the identical configuration used for collecting data on fundamental pitch. Recall that the sound levels were set to avoid clipping of the sound signal. The threshold was chosen so that all notes triggered the detector (*i.e.*, no soft notes fell below the threshold) and as many consonants as possible were excluded from processing. Each performance was played into the detector, and a sequence of spectral envelopes with time stamps was recorded to a file. The previously described method was used to generate codebooks containing from 1 to 256 entries. For automated speech recognition systems, codebooks of size 256 have been found to minimize quantization error sufficiently. For this project, the codebook with 128 entries was selected. The resulting quantization error was comparable to the codebook containing 256 entries, but the smaller number of entries resulted in smoother distributions using the comparably small amount of data generated from the recordings. 128 entries also provided a useful differentiation of the distributions for the different vowels.

To estimate distributions for each phoneme, each recording first was segmented by hand in order to determine the time of each distinct phoneme in the performance. The recordings were transferred to a computer representation and examined using an application for waveform visualization and playback. Phonetic transcriptions of the lyrics were generated for each score. Since the vocalists in the recordings had not been asked to provide phonetic transcriptions of their pieces, these transcriptions were obtained through consulting a combination of published transcriptions of songs (Coffin 1982), texts on diction, dictionaries for the various languages, and in some instances carefully listening to the actual recordings. Note that the objective was to identify the intended diction, not the actual diction used, which might have varied from the score. The start time of each phoneme in the transcription was obtained by using the waveform visualization and playback tool. Consistent, steady-state portions of phonemes are easily determined. While it is impossible to locate (or perhaps even to define) the exact point of change from one phoneme to another, a transition region of less than 100 ms can always be identified. In most cases, a clear transition region less than 50 ms is observed. By selecting the center of this region as the start of the later phoneme, the error of the start time estimate is always guaranteed below 50 ms, and is most often guaranteed below 25 ms. In addition, to the extent that one believes it is plausible to attribute the first half of the transition region to the earlier phoneme and the second half to the later phoneme, the actual error in all cases is below 25 ms.

The phonetic transcription with start times was aligned with the time-stamped sequence of spectral envelopes for each performance. All observed spectral envelopes that fell within the time spanned by each phoneme were attributed to that phoneme. The representation of spectral envelope was quantized using the previously described vector quantization codebook and Euclidean distance metric. For each IPA symbol, a distribution over the vector codebook was developed by simply counting the frequency with which each codebook entry was attributed to that symbol in the transcriptions. The number of observations per symbol ranged from a high of 1635 to a low of 4. As expected, a low number of observations was associated with many consonants, but even several vowels and diphthongs were associated with fewer than 100 observations. These phonemes were sounds unique to either English or German. Such low numbers of observations cannot provide accurate approximations to the actual distributions for the respective phonemes.

Because of the lack of data, the distributions for individual phonemes had to be combined, thereby forming distributions for equivalence classes of phonemes. A method for determining shared distributions for multiple phonemes was applied. Combining statistical models and distributions is commonly required when constructing automated speech recognition systems. Phoneme distributions are combined based upon a measure of difference between the distributions. Several methods have been proposed, including merging models based on knowledge of phoneme production (Derouault 1987) and measures derived from statistical assumptions and information theory (Paul and Martin 1988; Lee 1990). Ideally, one would like to combine the models for data sets whose true underlying distributions are most similar. Unfortunately, the statistical assumptions or objectives of most merging techniques are invalidated when using small samples to estimate the relatively arbitrary distributions over a vector codebook. Consequently, the difference metric actually applied constituted a simple sum over the difference in probability for each codebook entry:

$$\mathit{Difference}(\mathit{Phoneme}1, \mathit{Phoneme}2) \quad = \sum_{v \in \mathit{Codebook}} \left| \, \mathrm{P}[v|\mathit{Phoneme}1] - \mathrm{P}[v]\mathit{Phoneme}2] \, \right|$$

This calculation ranges from a value of zero when the estimated phoneme distributions are identical to a value of two when the distributions do not overlap. Phonemes were combined by repeatedly merging the distribution estimated from the smallest sample until no remaining distributions were estimated from fewer than 200 observations. Each distribution estimated from the smallest remaining sample is merged with the distribution that minimizes the given difference measure. Note that the difference measure calculates the area of the nonoverlapping region of two probability functions. Intuitively, this process merges one phoneme with another based upon how closely the distribution of actual observations for the latter phoneme match as the distribution of actual observations for the former. Such merging implies that we can do no better than to account as best as possible for the distribution obtained from the few observations we have seen.

Figures 5-6 and 5-7 depict the distributions over the vector codebook that were generated by combining per phoneme distributions as described. Note that many of the equivalence classes seem intuitively reasonable. Also, fourteen of the fricatives, affricates, and stops have been grouped together into one class. Individual distributions for these consonants contained relatively few observations over similar codebook entries. Perhaps the most astonishing class contains the vowel [i] and the stop [p]. The commonality of observations between these two phonemes might result from a common tongue position when singing [p] after a vowel or during the transition into a vowel following [p]. Possibly the observations from the recorded performances contain examples of loud [p]'s following or preceding a high front vowel like [i]. Nevertheless, the equivalence classes overall appear to be reasonable.

The distributions over the vector codebook entries do offer some discrimination between vowels. Note for instance that the distribution for [i] assigns high probabilities in the regions spanning codebook entries 48 to 64 and 80 to 90, while the distributions for the back vowels [o] and [oʊ] assign high probabilities in the regions spanning entries 8 to 32, 64 to 75, and 100 to 112. In contrast, the vowels [i] and [I] are known to have similar formants, so it is not surprising to see a similarity between the distributions for these vowels. Such comparisons and apparent distinctions are encouraging. Since it is not possible to make strong statements about proximity of consecutive codebook entries within the vector space, assessing the degree of smoothness and the variance of the graphed distributions is not helpful. Likewise, it is not useful to assess unimodality or normality of these graphs. Recall that these considerations were helpful in assessing accuracy of the distribution for observed fundamental pitch—a distribution also over a discrete dimension, but one ordered in a physically (and musically) meaningful way.

One useful assessment is to consider the likelihood of confusing one phoneme with another based upon a single observation of spectral envelope. Recall this calculation was presented when assessing the distribution over semitone difference between observed and scored pitch. It assumes that the estimated distributions adequately represent both the true distribution of the observations and the distribution used within the score-following model. Table 5-3 presents the estimated likelihood of confusing one vowel with another for the seven vowels in Italian. Note that the vowels are sequenced according to tongue position, starting with the high front vowel [i] moving through the low vowel [ɑ] and ending with the high back vowel [u]. This ordering places vowels with similar formants in close proximity. Consequently, it is encouraging to see that the likelihood of confusion is highest near the

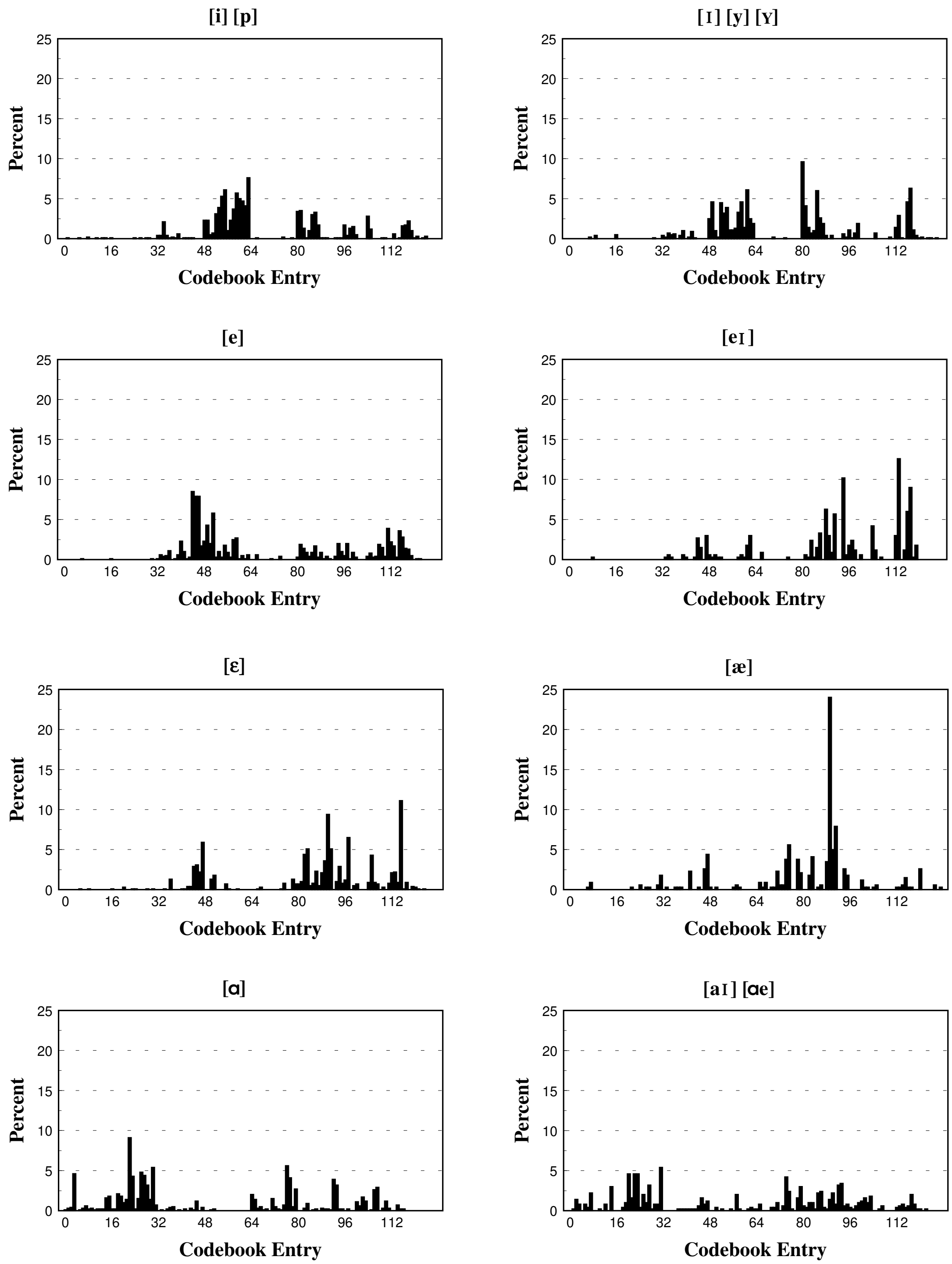


Figure 5-6. Distributions over the vector quantization codebook entries for phoneme equivalence classes containing all vowels that were modeled.

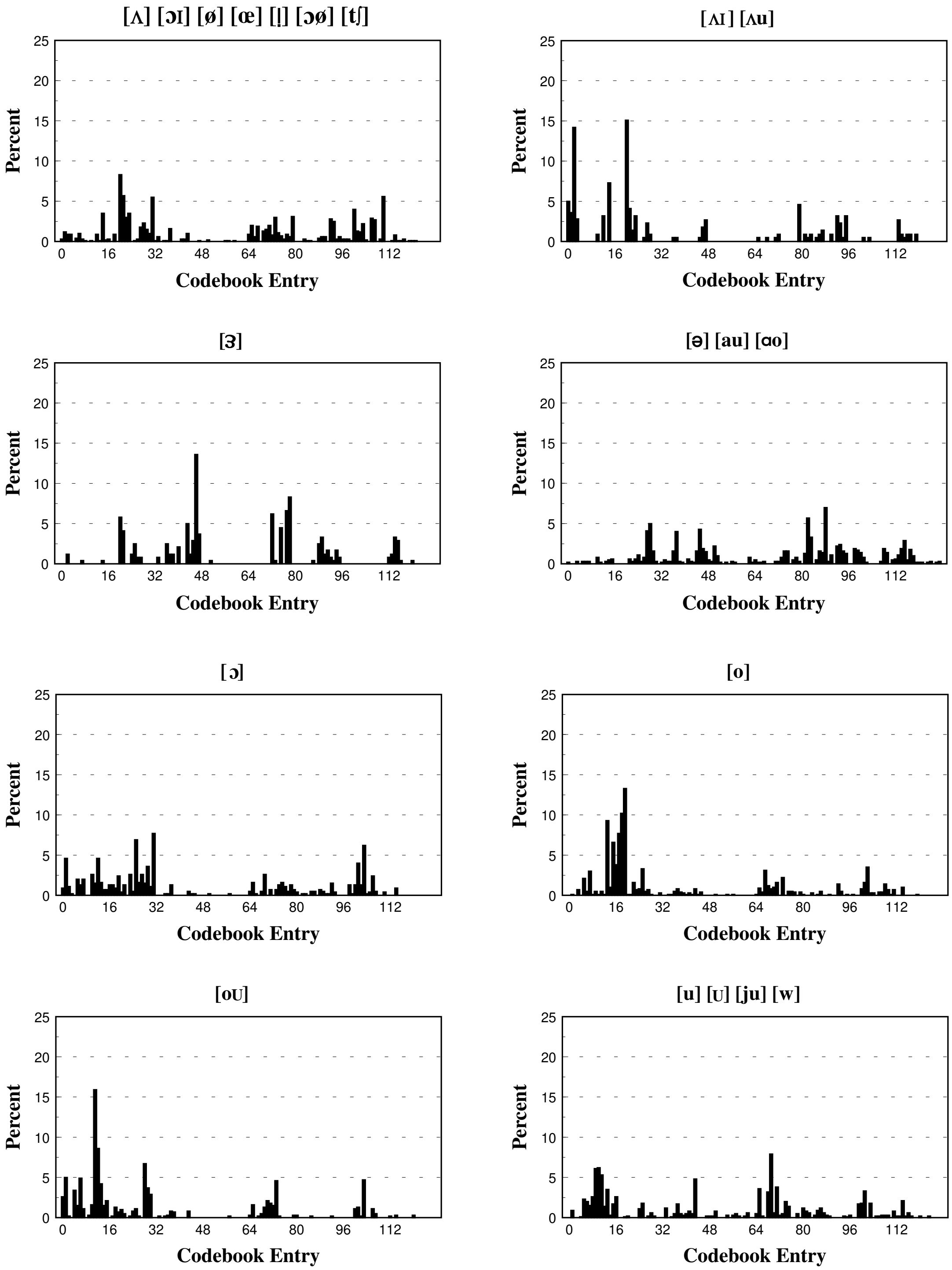


Figure 5-6. (cont.) Distributions over the vector quantization codebook entries for phoneme equivalence classes containing all vowels that were modeled.

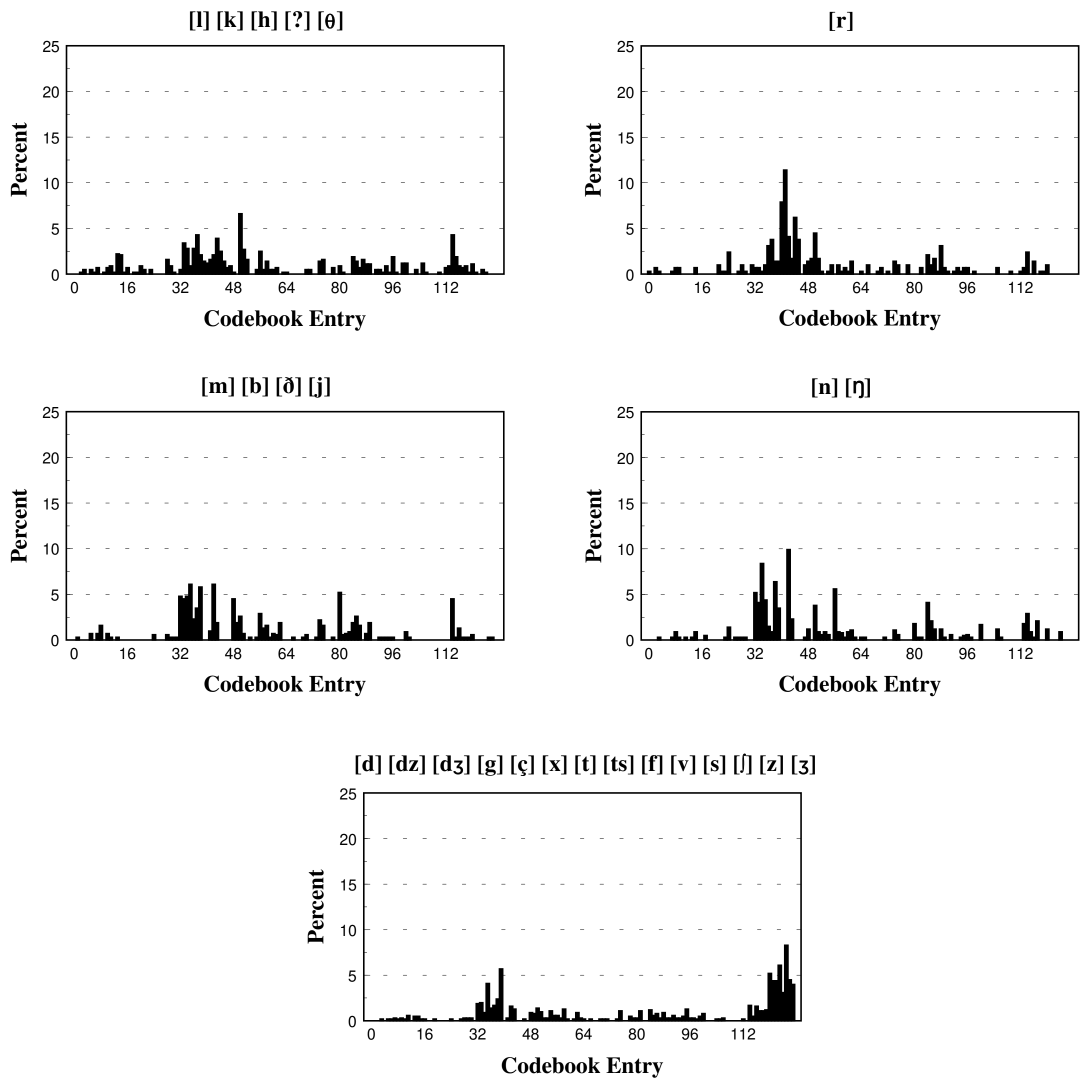


Figure 5-7. Distributions over the vector quantization codebook entries for phoneme equivalence classes containing only consonants.

diagonal. Also note the comparatively high values in the upper left and lower right quadrants of the table, indicating that overall the front vowels are more likely to be confused with other front vowels and the back vowels with other back vowels. Table 5-4 shows the likelihood of confusion for the ten distributions over the English monophthongs. The vowels are sequenced from the front vowels (high to low) to the mid vowels to the back vowels (low to high). Note that again high likelihood of confusion generally occurs near the diagonal, spreading out around the mid vowels. In particular, the neutral vowel [ə] achieves a double digit likelihood of being confused with every other monophthong. Also note the high likelihood of confusing the vowels [i] as in *beet* and [ɪ] as in *bit*, and the vowels [ʌ] as in *shut* and [ɑ] as in *shot*.

Table 5-3. Estimated likelihood (percent) of confusing the seven vowels in Italian based on a single observation. Table entries indicate the likelihood that the model will assign an equal or higher probability to the vowel in the column when observations are generated from the vowel in the row.

	[i]	[e]	[ɛ]	[ɑ]	[ɔ]	[o]	[u]
[i]	100	12	7	3	4	3	5
[e]	26	100	22	3	3	2	6
[ɛ]	11	20	100	6	6	3	8
[ɑ]	2	7	8	100	17	18	16
[ɔ]	4	5	8	25	100	20	25
[o]	2	7	6	14	17	100	19
[u]	10	12	8	6	16	15	100

The observed confusability of certain pairs or groups of modeled vowels can be attributed to several factors. First, as indicated by the ordering of vowels in the tables, certain pairs of vowels simply have similar spectra. Some overlap in the distributions for these vowels should be expected. Also, it is well known that certain vowel pairs sound similar when sung on a high fundamental pitch. These pairs include [i]-[ɪ] and [ʌ]-[ɑ]. The vowels in such pairs have similar formants, so it becomes difficult to distinguish the spectra of these vowels when the harmonics are sparse. Second, combining distributions for individual phonemes can result in a final distribution that results in a higher likelihood of confusion

Table 5-4. Estimated likelihood (percent) of confusing the monophthongs in English based on a single observation. Table entries indicate the likelihood that the model will assign an equal or higher probability to the vowel in the column when observations are generated from the vowel in the row.

	[i]	[ɪ]	[ɛ]	[æ]	[ɜ]	[ə]	[ʌ]	[ɑ]	[ɔ]	[u]
[i]	100	36	7	3	3	11	3	3	4	5
[ɪ]	33	100	9	4	3	7	2	2	2	2
[ɛ]	11	11	100	18	10	21	6	6	6	8
[æ]	4	5	23	100	13	16	12	6	9	10
[ɜ]	1	2	22	21	100	16	13	3	8	3
[ə]	10	16	22	19	22	100	15	17	14	18
[ʌ]	3	4	8	9	10	16	100	27	27	18
[ɑ]	2	3	8	11	16	16	26	100	17	16
[ɔ]	4	3	8	10	10	19	22	25	100	25
[u]	10	14	8	9	16	17	10	6	16	100

for more vowels. For instance, the distribution used for [u] is estimated from observations for several phonemes including [u], [U], [ju], and [w]. It is not surprising that this distribution shows double digit likelihood of being confused with vowels similar to [u], [U], and [j]. A similar situation should be expected for explicitly modeled diphthongs as well. Finally, similarities in the distributions can result from inaccuracies in the phonetic transcription, the parsing of the recordings, or the time alignment of the parsed recording and detector output. For instance, a singer may have intentionally sung [e] as in the Italian *che* where an [ɛ] as in *bed* was transcribed.

Based on the likelihood of confusing vowels when considering a single observation, it is evident that spectral envelope is not an ideal type of observation. The major drawback to using spectral envelope is the large number of variables needed to specify a truly comprehensive distribution function. Not only does the observation itself consist of multiple values, but several important factors affect the observation value. Such a large number of variables prevents collection of sufficient vocal performances to adequately define arbitrary distribution models. Unlike the distance distribution modeled in Chapter 3, no analytical model relating factors and observations is readily apparent, so appropriate methods for smoothing or interpolating with the available data have not been determined. Nevertheless, the distributions estimated for the automatically generated phoneme classes are used to specify the distribution of observed spectral envelope for events consisting of phonemes appearing in a transcription of the lyrics.

One final adjustment was made to the graphed distributions prior to using them for actual tracking. Since the samples used to estimate the distributions ranged in size from under 300 observations to more than 1500, some distributions are more sensitive to infrequently appearing codebook entries. For instance, a codebook entry that appears with an actual probability of 0.001 (on average once for every 1000 observations) is more likely to be observed in a data set of 1500 observations than a data set with only 300. Just as with fundamental pitch, no observation distribution should assign a probability of 0 to any observable value unless it is truly impossible to observe that value. Even if this codebook entry does appear in a sample of 300 observations, it will be assigned a probability of at least 0.0033, while in the sample of 1500 the assigned probability could be as low as 0.00067. The estimated probabilities may differ by at least a factor of 5. Suppose that these two samples define distributions for two different phonemes, and the actual probability of observing this codebook entry is identical for both phonemes. The modeled distributions will indicate erroneously that the observation is five times more likely for the phoneme modeled using a smaller sample. Although this observation appears infrequently when either phoneme is performed, it can have a serious effect on position estimation when it does surface.

To minimize the effects of modeling phonemes with different size samples, a minimum probability of observing any possible vector codebook entry for any phoneme is determined. Since the smallest sample contained around 300 observations, this minimum is set at a value of 0.003. All probabilities for all codebook entries over all phonemes are assigned a value at least this large. Imposing this limit corresponds to assuming equal likelihood for all codebook entries that appeared with a probability below the smallest nonzero probability that the smallest sample distinguishes. Thus when the probability of observing a given codebook entry falls below this sensitivity level for each of two distinct phonemes, the stochastic score-tracking model will not prefer either phoneme based solely on observing that codebook entry. The model ignores certain distinctions that probably are inaccurate due to a known limitation of distribution estimation.

The final problem for modeling events based on phonemes is determining the score positions spanned by each phoneme occurring in the transcription of the lyrics. Phoneme positions in turn depend upon the duration of each phoneme. While a length for each note in the score can be generated using a fixed, idealized tempo as in Chapter 4, determining length of individual phonemes for each note requires additional information. For notes that are not too short, the vowel will span the largest portion of the note length. Consonants in the syllable span a noticeably shorter length. Table 5-5 contains the average durations of consonants appearing at least 50 times in the set of recorded performances. The average over all consonants appearing in the recordings (more than 2500 in total) was 115 ms. Since in some recordings it was difficult to determine the start of a consonant that follows a rest or the end of a consonant that precedes a rest, the duration of individual consonants in these positions were probably overestimated. Consequently, the average durations shown in the table are probably larger than the true averages. However, the average duration of 115 ms was used as the model length for all consonants in the phonetic transcription of the lyrics.

The starting position of each phoneme is determined according to the following simple rules. The length of each note in the score is calculated as for the events based on fundamental pitch. Namely, the relative duration of each note is multiplied by a nominal tempo. Recall that this tempo is selected to be a minimum expected tempo for a piece or a section of a piece. A minimum is used in order to avoid undersampling the distance density which can lead to poor approximation of the convolution integral. The vowel in each syllable is assumed to start at the beginning of the note. In the case of a textual diphthong for which no explicit model was generated, the primary vowel of the diphthong (*i.e.*, the vowel expected to be longer) is assumed to align with the beginning of the note. The secondary vowel is treated as a consonant. All consonants appearing between two vowels are assumed to span portions of the preceding note, thereby reducing the expected duration of the first vowel. Consonants between a rest and a vowel, or a vowel and a rest, are assumed to span portions of the rest. Consonants and secondary

Table 5-5. Average duration of individual phonemes that appeared at least 50 times in the recorded performances, and the average duration over all phonemes in the performances (excluding vowels).

Phoneme	Mean Duration	Phoneme	Mean Duration
[b]	92 ms	[r]	113 ms
[d]	78 ms	[s]	155 ms
[f]	118 ms	[ʃ]	163 ms
[h]	100 ms	[t]	100 ms
[k]	102 ms	[v]	106 ms
[l]	140 ms	[w]	146 ms
[m]	122 ms	[y]	113 ms
[n]	113 ms	[z]	116 ms
[p]	103 ms	All Phonemes	115 ms

vowels in diphthongs are assumed to have a length of 115 ms each. The only exception is when the sum of lengths for the consonants would reduce the length of a vowel below 115 ms. In this case, the length of the note is apportioned evenly amongst the vowel and consonants.

Although not entirely accurate, the given method for determining position of each phoneme is not unreasonable. First, singers are often encouraged to begin the vowel at the start of the note. In reality this does not always happen, especially if the immediately preceding consonant is a strongly voiced consonant such as [l], [m], or [w]. However, no data was collected in order to examine a singer's alignment of phonemes with the accompaniment, so the simple assumption that the vowel is aligned with the start of the note was applied consistently. Second, the duration of consonants is certainly not increased in direct proportion with increases in vowel duration. After some minimum note duration is surpassed, increases in note duration correspond solely to increases in vowel duration. However, it is also true that in cases where there are "too many" consonants relative to the total note duration, singers will shorten the duration of both consonants and vowel so that all phonemes "fit into" the note durations. While no empirically based model was developed to describe the lengthening and shortening of consonants, the selected method of specifying phoneme lengths does exhibit this type of behavior. Finally, although tempo changes made by the singer effectively cause the score-tracking model to scale all lengths proportionately, the relatively small length of the modeled consonants (115 ms) is not likely to translate into significant absolute increases in duration. Tempo changes of 10 to 20 percent are very significant, and this translates into less than 25 ms for the modeled consonants. The resulting scaled durations are not far from the estimated average durations for some consonants.

5.5 Summary of Observations Based on Spectral Envelope

The stochastic method for score following incorporates a probability density defining the likelihood of all possible observation values at any score position. This chapter has considered the use of spectral envelope as one type of observation, either in isolation or in combination with fundamental pitch. Several problems with predicting the spectral envelope of sung phonemes were noted. Representations of the spectral envelope were discussed, and the details of an envelope detector implemented for this project were presented. This detector was applied to a set of eighteen vocal performances. Using the output of the detector, an approximation to the observation density function, $f_{V|I}(v \mid i)$, for spectral envelope was presented in the previous section. This approximation estimates the likelihood of an observed spectral envelope conditioned on a phonetic transcription of the score. Although the observed spectral envelope is not independent of the fundamental pitch, combining observations of spectral envelope and fundamental pitch is done by assuming the observation types are independent and reporting both types of observations in synchrony. This approach permits use of the following approximation to the observation density:

$$f_{Pitch,Envelope|I} \cong f_{Pitch|ScoredPitch(I)} \cdot f_{Envelope|ScoredPhoneme(I)}$$

The accuracy of the stochastic score-following model when assuming this observation density was empirically evaluated, and is discussed in a subsequent chapter.

The adequacy of the estimated distributions for observed spectral envelope is more questionable than for any distribution so far presented. In particular, the lack of data to sufficiently estimate the arbitrary distribution models is worrisome. Lack of an analytical model for relating the relevant and numerous factors to the vector extracted from the log spectrum is also cause for concern. Furthermore, these two problems conspire to prevent inclusion of all obviously relevant factors in the distribution model, including fundamental pitch. If the estimated distributions over spectral envelope are poor approximations to the actual, then the approximation to the combined distribution over pitch and envelope may be inadequate as well. However, since the lack of data is a significant problem, direct estimation of the joint density over pitch and envelope probably would not improve the approximation. In fact, to the extent that the distribution over fundamental pitch is well approximated and includes the most relevant factors, the approximation to the joint density based on the independence assumption may actually be preferable to estimates using only a handful of data points for each possible combination over all the conditioning variables. An insightful decomposition of the joint distribution can be helpful.

The complete model for the stochastic score-following system based on observations of fundamental pitch and spectral envelope is presented in Figure 5-8. The model incorporates distributions for observations described in this chapter and the previous chapter. Note that the model can be altered to work with only one type of observation by setting the density for the other observation type equal to a uniform density. One question of interest is whether or not the extended model including observations of spectral envelope provides better performance tracking than the model based only on observing fundamental pitch. Considering the difficulties in accurately modeling distributions for observed spectral envelope, it is not intuitively obvious that combining one imperfect observation with another, probably less perfect observation will yield a more accurate position estimate. Although final comparative assessment of the two models requires empirical evaluation, an intuitive explanation will be provided for why improvements are still possible when using multiple observation types. However, thorough discussion of this statistical property is postponed until the end of the subsequent chapter.

The next chapter contains a presentation of the third and final type of observation that has been incorporated into the stochastic score-following model. This observation is based on measuring changes in amplitude that indicate the singer has transitioned to a new note in the score. Subsequent chapters describe the results of applying various versions of the score-following model to accompany both recorded and live performances by vocalists.

Vocal Performance Tracking Model:

$$f_I^{t1}(i) \;=\; f_{I\,|\,J,\,\mathrm{D},\,V}\,(i|\; j = j_0,\; d = d_0,\; v = v_0[Pitch, Envelope])$$

$$\cong \; \frac{f_{V\,|\,I}\,(v = v_0[Pitch, Envelope]|\; i) \cdot C(i,j,d = d_0)}{\int_{i=0}^{\|Score\|} f_{V\,|\,I}\,(v = v_0[Pitch, Envelope]|\; i) \cdot C(i,j,d = d_0)\;\partial i}$$

where $\quad C(i,j,d = d_0) \;\cong\; \int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\; d = d_0) \cdot f_I^{t0}(j)\;\partial j$

j = source position of the performer.

i = destination position of the performer.

d = estimated score distance traversed by the performer.

v = newly reported observation.

Distribution of Actual Score Performed:

$$f_{I-J\,|\,D}(i-j|\; d = d_0) \;\cong\; f_{I-J\,|\,R,\,\Delta T}\,(i-j|\; r = r_0,\; \Delta t = \Delta t_0) \;\cong\; \frac{1}{(i-j)\sigma\sqrt{2\pi}}\; e^{\frac{-(\ln(i-j)-\mu)^2}{2\sigma^2}}$$

where $\quad \mu = \ln r_0 - \frac{1}{2}\sigma^2 + \ln \Delta t_0$

$\sigma^2 = \ln\left(\frac{1}{.02948\Delta t_0} + 1\right)$

$\Delta t_0 = t1 - t0$

r_0 = estimated average rate over preceeding ~3 seconds.

Distribution of Observations:

$$f_{V\,|\,\mathrm{I}}(v = v_0[Pitch, Envelope]|\; i) \;\cong$$

$$f_{V\,|\,\mathrm{ScoredPitch}}\,(v = v_0[Pitch]|\, ScoredPitch(i)) \;\cdot\; f_{V\,|\,\mathrm{ScoredPhoneme}}\,(v = v_0[Envelope]|\, ScoredPhoneme(i))$$

Distribution of Observed Fundamental Pitch:

$$f_{V\,|\,\mathrm{ScoredPitch}}\,(v = v_0[Pitch]|\; ScoredPitch(i)) \;\propto\; \mathrm{P}_{SemitoneOffset}[v_0[Pitch] - ScoredPitch(i)]$$

$$\cong \text{ probabilities from Table 4-1}$$

Distribution of Observed Spectral Envelope:

$$f_{V\,|\,\mathrm{ScoredPhoneme}}\,(v = v_0[Envelope]|\; ScoredPhoneme(i)) \;\cong\; \text{distributions in Figures 5-6 and 5-7}$$

Figure 5-8. Stochastic score-following model using fundamental pitch, spectral envelope, rate, elapsed time, and source position.

Chapter 6

Events Based on Observing Note Onsets

6.1 Measuring Changes in Amplitude and Detecting Note Onsets

A complete stochastic model for tracking vocal performances was presented in Chapter 4. This model includes estimated distributions for both the actual amount of score performed between observations and detected fundamental pitch. The model was extended in Chapter 5 to include observations of spectral envelope. The enhanced model incorporates both types of observations by assuming independence. The joint distribution of the observations conditioned on score position is approximated by a product of the individual density functions for each type of observation:

$$f_{Pitch,Envelope|I} \cong f_{Pitch|ScoredPitch(I)} \cdot f_{Envelope|ScoredPhoneme(I)}$$

In this chapter, the score-tracking model is extended to include a third type of observation—amplitude changes in the sound signal that indicate the beginning of a note, or the *note onset*.

To include detected note onsets in the score-following model of Chapter 5, the observation density must be redefined as a joint distribution over three observation types. Either the joint distribution must be specified explicitly or an approximation based on one or more simplifying assumptions must be used. The model presented in this chapter takes the second approach. Specifically, it assumes that all three observations are independent of one another:

$$f_{Pitch,Envelope,Onset|I} = f_{Pitch|I} \cdot f_{Envelope|I,Pitch} \cdot f_{Onset|I,Pitch,Envelope} \cong f_{Pitch|I} \cdot f_{Envelope|I} \cdot f_{Onset|I}$$

Assuming independence between pitch and spectral envelope was discussed previously in Chapter 5. Similarly, the definition of an observed onset and the conditioning variables (important factors) influencing onset detection ideally should support the independence assumptions. As with the other observation types considered, observed note onsets must be independent of the performer's previous score position and the estimated distance (*i.e.*, estimated tempo and elapsed time). Recall that these independence assumptions are used to derive the implemented score-tracking model from the general model. Also, reporting of note onsets is synchronized with reports of fundamental pitch and spectral envelope. As already described, synchronized observations are necessary to avoid unacceptable numerical error when approximating integrals in the score-tracking model. Note that this limitation is due to the real-time implementation of the model and is not an assumption of the theoretical, continuous model.

Note onsets are an important indicator of score position, especially when the ultimate goal is to synchronize different parts in a performance. In musical scores, the notes in different parts often align. Since the onset indicates a significant change in the signal, misaligned onsets between performers can be more noticeable than timing alterations made by one performer while another performer sustains a note. Accurate identification of onsets is therefore important in musical tracking tasks. Depending upon the signal analysis performed for onset detection, reported onsets will correspond more or less identically with the start of each new note. Methods for detecting onsets actually indicate points of significant change in the performance signal. In this respect, onsets differ from the previously considered observation types of fundamental pitch and spectral envelope that characterize steady-state features of the signal.

Several methods for detecting note onsets have been reported. They are applied both for score-following applications and for automated music transcription software. The simplest methods are based on detecting significant changes in the amplitude of the signal. Generally the change in amplitude or power over a restricted interval of time (*i.e.*, energy of the signal) must be considered. Selection of the time interval is important. If the time interval is too small, amplitude changes due to periodicity of the signal can trigger false detection. Conversely, if the time interval is too large, short transitions between notes will fail to trigger onset detection. The criteria for distinguishing significant amplitude changes from insignificant changes are also important. Examples include recognizing a drop below a preset threshold followed by a rise above that threshold or identifying relative changes exceeding a preset, minimum factor (Puckette 1995).

Foster, Schloss and Rockmore (1982) describe methods of onset detection used in a music transcription system. In addition to changes in amplitude, they consider both identifying changes in fundamental pitch and the technique of autoregressive segmentation. For many instruments, detected pitch and signal intensity remain fairly consistent during the steady-state portion of a note but change noticeably around the onset. In such cases, reliable onset detection is possible by clustering short time segments of a performance signal based on pitch and intensity. Since this technique relies on consistency of pitch and long time clustering, it would be less effective for real-time segmentation of vocal performances. Autoregressive segmentation fits an autoregressive model to two consecutive, short time segments of signal and compares how well each model fits the most recent segment. Recall that an nth-order autoregression estimates each sample by a linear combination of the n immediately preceding samples plus an error term:

$$s[j] = \sum_{i=1}^{n} a_i \, s[j-i] + \varepsilon_j$$

The segmentation algorithm computes the average squared error term, ε_j^2, when using each set of autoregressive coefficients to predict the most recent segment of signal. In a steady-state region of signal these error terms will be comparable, so the ratio of the errors will be close to one. When the two segments lie on either side of the boundary between different steady-state regions, the error term for the model generated from the first segment will be greater. Thus continuously sliding a window through the signal and computing the ratio of the squared error terms will generate a function with noticeable peaks at the boundary between different steady-state regions of the signal. To the extent that the autoregressive model adequately characterizes the musical signal, these boundaries often correspond to note onsets.

Another approach to identifying onsets considers differences between the spectra of consecutive, short time segments of signal. Since the spectrum remains relatively consistent within a steady-state region of the signal, large differences in the spectrum often will indicate a note onset. This approach is similar in spirit to the autoregressive segmentation just described. Some recent work in automated music segmentation has considered the difference between "spectra" (more precisely, scale or modulus planes) computed using wavelet transforms (Solbach and Wöhrmann 1996; Tait and Findlay 1996). Similarly, systems for automated speech recognition often use differences between two successive spectra or cepstral vectors each computed from short segments of signal (Furui 1986; Gupta, Lenning and Mermelstein 1987; Huang *et al.* 1991; Lee *et al.* 1992). While the actual values of these spectral or cepstral differences are used directly for phoneme and word identification, they undoubtedly provide information about the boundaries between phonemes in the speech signal. It is reasonable to posit that these calculated differences are useful for word identification primarily because they distinguish "phoneme onsets" and the transitions between different pairs of phonemes.

Onset detection for this project is based on identifying significant changes in amplitude over a minimum elapsed time. Detection is based on the amplitude threshold used in detecting fundamental pitch and spectral envelope. This threshold is set so as to distinguish pitched signal from unpitched signal. However, transitions from amplitude below the threshold to amplitude above the threshold often indicate the start of a note. Sung vowels often exhibit a significant change in amplitude when preceded by consonants, a rest, or a breath. For this project, the onset detector reports an onset when 30 ms of signal with amplitude below the threshold precedes a detected pitch. 30 ms is more than twice the pitch period of the lowest note sung by a bass, but typically less than the shortest consonants appearing between vowels. Recall that the defined fundamental pitch detector requires 100 ms of signal above the threshold before reporting pitch. Detected onsets must be spaced by at least 150 ms, otherwise the detector does not report the second onset. This constraint reduces duplicate reports during note transitions. 150 ms is almost always shorter than the duration of sung notes not in a melisma. The detector reports presence or absence of an onset for every observation of fundamental pitch and spectral envelope. Providing that the amplitude threshold is reasonable, this simple approach can reliably detect a useful number of note onsets.

The defined onset detector applies a very simple approach to identify onsets and is subject to several limitations. First, it does not report the onset of every note, especially notes in a melisma where the amplitude does not fall below the threshold. In addition, a poor setting for the amplitude threshold can result in missed onsets or spurious detections due to loud consonants or dynamic changes on sustained notes. However, changes in amplitude are influenced by many factors. Applying a simple method of onset detection minimizes the number of factors that must be considered. As will be shown, the onsets reported by this approach are determined mainly by consonant to vowel transitions, rests and breaths. Thus there are relatively few conditioning variables to consider. In contrast, continuous measurements such as autocorrelation and spectral derivatives are influenced also by the specific phonemes and dynamic changes. Consequently, statistics for the reported onsets would be more challenging to model and require more data if autocorrelation or spectral derivatives were used.

6.2 Onset Detection for Vocal Performances

There are five criteria that an ideal observation type should satisfy. These five criteria consist of the following: position discrimination, the ease of specifying the distributions, the availability of significant factors affecting the observation, the ease of estimating the distributions, and consistency with the score-following model. Realistic types of observations inevitably fail to satisfy these requirements completely. As with fundamental pitch and spectral envelope, amplitude changes associated with note onsets are no exception. This observation type also fails to satisfy the requirements completely.

First, note onsets cannot completely discriminate score position. As defined, they at best can distinguish a region near the beginning of each note from the remainder of the note. On their own, they offer little ability to distinguish the beginnings of different notes in the score. Compared to fundamental pitch and spectral envelope, this limited discrimination ability seems impoverished. On the other hand, recognition of an onset often will distinguish a relatively small region at the beginning of the note. The distinction will be sharp and will complement the regions distinguished by observing pitch and envelope. Thus while note onsets do not offer the ideal type of score position identification, they are informative indicators of a specific situation (namely a performer starting the vowel) that supplements position identification using the previously discussed observation types.

As an observation type, note onsets do not permit easy specification of distributions for every point in the score. Several factors hinder precise definition of the distributions. First, as discussed in Chapter 5, the transitions between phonemes are not clearly indicated in a musical score. Since onset recognition often is triggered by a change in phoneme, lack of precise phoneme positioning complicates modeling of this observation type as well as spectral envelope. Second, observed note onsets do not indicate steady-state signal properties as do fundamental pitch and spectral envelope. Thus the true likelihood of observing an onset probably changes continuously within each note, even near the beginning of the note. This property calls into question the use of events to define sets of contiguous score positions that have identical observation probabilities.

In addition to the previously mentioned problems, modeling of onsets is further complicated by the numerous conditioning variables that influence onset recognition. First, the phonetic content of the score certainly influences recognition of onsets. Transitions from a sung consonant to a vowel will often produce significant changes in amplitude, but not always. Change in amplitude is often affected by the number of consonants prior to a vowel as well as whether or not the consonants are voiced or unvoiced. Second, recognition of onsets is also influenced by the location of rests and the singer's breathing. While many rests will be indicated explicitly in the musical score, the singer may on occasion add a short rest between notes, often for the purpose of breathing prior to singing the next phrase. Although trained singers generally plan the breaths in advance, there is no guarantee that such predetermined breathing will be observed consistently. Third, dynamic changes in the performance can produce significant changes in amplitude. For instance, a note that starts softly and crescendos as it is sustained can trigger onset recognition if the amplitude threshold has been set too high. Finally, position of the microphone relative to the singer will alter the amplitude of the processed signal. If the singer moves or changes head position during a performance, spurious onsets may be reported or actual onsets may be missed.

As with previously discussed observation types, collecting data for distribution estimation is tractable if it is reasonable to partition the score into events and only a limited number of conditioning variables participate in the model. However, neither of these requirements is satisfied by note onsets. As mentioned, the true likelihood of observing an onset changes continuously over the region of score spanned by each note. In addition, the numerous factors influencing the observed spectral envelope hinders data collection for this observation type. It is unlikely that a model including all the previously identified factors could be estimated adequately, especially considering the difficulty in collecting examples of operatic performances.

Finally, observations of spectral envelope may not entirely satisfy the assumptions of the score-following model. Although the chosen representation of a note onset can be reported as a fixed value, it is not valid to assume that these observations are entirely independent of the previous position of the performer and the estimated amount of score performed. Specifically, the tempo and previous position of the performer influence whether or not an onset will be observed at any particular point near the beginning of the note. Thus incorporating rate, elapsed time, and previous position as conditioning variables in the distribution for note onsets would lead to improved estimates, providing that a sufficient number of example performances could be collected. The influence of rate and previous position may be more important for observation of note onsets than for either fundamental pitch or spectral envelope.

Just like observations based on fundamental pitch or spectral envelope, observations based on note onsets fail to completely satisfy the criteria for an ideal type of observation. While each type of observation satisfies the criteria to different degrees, none is sufficient to guarantee successful vocal performance tracking if used alone. However, using all three observation types in combination can lead to improved tracking, even when the statistical models and estimated values are imperfect. In the next section of this chapter, a model of detected note onsets (significant changes in amplitude) is presented. Subsequent to defining this model, some theoretical motivation is provided for expecting improved tracking of vocal performance when combining observations of fundamental pitch, spectral envelope, and note onsets.

6.3 A Model of Detected Note Onsets

The stochastic score-following model includes a function, $f_{V|I}(v \mid i)$, that specifies the likelihood of making a particular observation given the current score position of the performer. The first section of this chapter specified a decomposition of this function as a product of three distinct density functions, each density characterizing a different type of observation. This section describes an approximation to the density function for observations of amplitude changes indicating note onsets. Ideally, this approximation should include all primary factors influencing detection of onsets. Numerous such factors have already been identified. However, due to the limited amount of data available for estimating the density function, the factors actually considered have been restricted to a small subset of all possible factors.

The actual estimated distribution for detected onsets contains two important conditioning variables. First, detection of an onset almost always occurs during or just after note transitions. Onsets are almost never reported during a sustained tone. The likelihood of detecting the onset of a note depends upon whether the vowel immediately follows a rest or breath, consonants, or another vowel. In addition, onsets are sometimes detected immediately before a rest. Second, when a group of consonants precedes a vowel, the number of voiced consonants in that group affects the likelihood of detecting an onset. The presence of fewer voiced consonants increases the likelihood that an onset will be observed. Considering these factors, the estimated distribution for detected onsets specifies the likelihood of detecting onsets according to the following conditions: whether a transition between notes contains a rest or breath, consonants, or neither; and when consonants are present, both the number of consonants and the number of voiced consonants present. Using this distribution corresponds to the following approximation:

$$f_{Onset|I} \cong f_{Onset|TransitionClass(I),VoicedConsonants(I)} \propto \mathrm{P}\left[Onset|TransitionClass(I), VoicedConsonants(I)\right]$$

where *TransitionClass(I)* is either rest, vowel, one consonant, two consonants, three or more consonants, middle of note, or end of note. *VoicedConsonants(I)* evaluates to zero, one, two, or three (for three or more voiced consonants) when the transition class indicates consonants are present, and zero for any other transition class. Note that more than three consonants can be sung between two vowels, as is the case when two words such as "and stars" are sung in sequence. The consonants [n], [d], [s] and [t] appear before the second vowel. Also, sometimes three or more of the consonants are voiced. Consider for instance the two words "and slowly" where [n], [d] and [l] appear before the vowel [oU].

The onset detector was applied to eighteen recorded vocal performances—the same performances used to estimate density functions for fundamental pitch and spectral envelope. These performances included two recordings by each of nine singers spanning all primary voice types. The pieces performed included a variety of styles, genres, and dynamic levels. Note that during collection of all three types of observations, the amplitude threshold was held consistent per recording. The playback levels and sound card levels were set so as to avoid any clipping of the sound signal while maximizing the dynamic range of the signal. Threshold levels were determined by running each recording through the detector a few times and observing whether some note onsets were not detected and whether spurious detections occurred. The threshold was set just high enough to prevent any spurious detections during soft but sustained notes in the performance. These settings were low enough to detect many note transitions in each recording. Once an acceptable threshold was determined, the recording was played into the detector and all detected onsets were recorded along with a time stamp.

Detected onsets were time-aligned against the phonetic segmentations generated for spectral envelope estimation. Each detected onset was associated with the performed note whose start time (*i.e.*, the estimated start of the vowel or pitch) was closest to the onset. Probabilities were estimated over all possible value assignments for the conditioning variables in the density function. These probabilities were determined by calculating the percent of scored notes for which an onset was detected. Thus, if the combined scores contained 20 note transitions in the transition class "Vowel" and 5 onsets were detected over all these instances, the estimated probability of observing an onset given the transition class "Vowel" would be 0.25. Note that for each value assignment, P[*No Onset*] = 1 – P[*Onset*].

The onset probabilities estimated from the 18 recorded performances are presented in Table 6-1. The performances contained a total of 2246 notes. For the transition class "3+ Consonants", the probabilities for 0 and 1 voiced consonants were estimated from a combined sample of 36 notes. Sample sizes ranged from 36 notes to 537 notes. The transition class "Rest" included cases where breaths had been marked in the scores. The recorded performances contained only one instance where an onset was detected in the middle of a note. This occurred because the singer breathed during a long, sustained note. Thus, observing an onset within the middle of a note is extremely unlikely but not impossible, so the probability for this case is set to 0.0001.

The probabilities from Table 6-1 are graphed in Figure 6-1. For each consonant transition class, this graph clearly shows a decrease in probability as the proportion of voiced consonants increases. The sharpest drop occurs between the cases for one unvoiced consonant and no unvoiced consonants. Also, as the number of consonants increases, the likelihood of detecting an onset becomes closer to the probability for the "Rest" transition class. Not surprisingly, the probability of observing an onset appears positively correlated with the opportunities for low amplitude signal to appear between notes. However, notice that the probability for the transition class "Rest" does not equal one. Sometimes a singer will hold a note through a short rest or will fail to observe a breath marking in the score, thus not triggering detection of an onset. Appropriately dividing the "Rest" class into two categories based on the length of the scored rest might yield at least one transition class that is always associated with an observed onset.

Table 6-1. Probability of observing an onset based on the transition class and the number of voiced consonants. Probabilities were estimated from the set of 18 recorded vocal performances. For the transition class "3+ Consonants", the probabilities for 0 and 1 voiced consonants were estimated from a combined sample.

	Number of Voiced Consonants			
Transition Class	0	1	2	3+
Vowel	0.0343	——	——	——
1 Consonant	0.9041	0.4022	——	——
2 Consonants	0.9167	0.8852	0.4708	——
3+ Consonants	0.9677	0.9677	0.8873	0.6744
Rest	0.9641	——	——	——
End of Note	0.0590	——	——	——

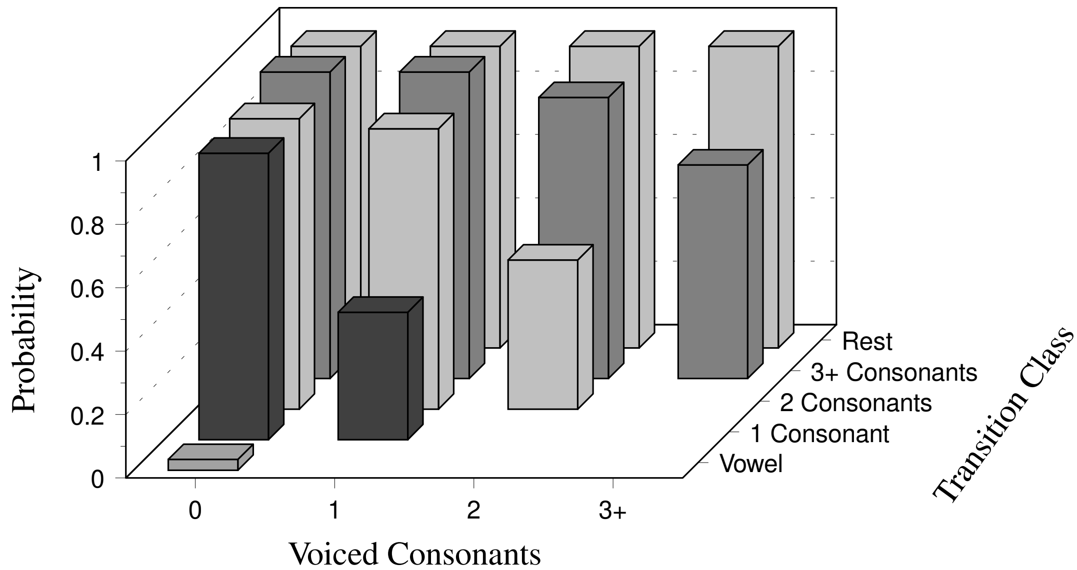


Figure 6-1. Probability of observing an onset based on the transition class and the number of voiced consonants (taken from Table 6-1). Probabilities were estimated from the set of 18 recorded vocal performances.

The entries in Table 6-1 define the density function for observation of events based on detecting note onsets. The accuracy of these estimates can be assessed more easily than in the case of the distributions estimated for fundamental pitch and spectral envelope. For each pair of value assignments to the conditioning variables, the onset observation distribution specifies a *Bernoulli random variable*. A Bernoulli random variable is a random variable that indicates only a success or failure condition (*i.e.*, it assumes a value of 0 or 1). The distribution for such a variable specifies the probability of success. The maximum likelihood estimate for the probability of success, p, is given by the mean of the sample, $\hat{p}$. The estimate of the variance is given by $\hat{p}(1-\hat{p})^2 + (1-\hat{p})(0-\hat{p})^2 = \hat{p}(1-\hat{p})$. Since $\hat{p}$ is a mean of a sample, it is asymptotically normally distributed and confidence intervals can be constructed for reasonable sample sizes using the standard error, the square root of $\hat{p}(1-\hat{p})/n$, where n indicates the size of the sample. Note that when estimating Bernoulli distributions, estimates producing extreme probabilities have smaller standard errors given a fixed number of data points. For instance, if estimating a Bernoulli distribution from a fixed size sample, the 95% confidence interval is smaller when a value of 1.0 results than when a value of 0.5 is produced.

Ideally, all the probabilities estimated for onset events should be near the extremes, indicating that the conditioning variables almost completely explain the detection of onsets. Unfortunately, the estimated probabilities are closer to 0.5 for the consonant transition classes when all consonants are voiced. The largest standard error for any estimate is around 0.05 and occurred for the case of 3 or more consonants, 3 or more voiced. The standard error for all other probabilities in the consonant transition classes were between .02 and .04. The standard error for rests and vowels were both very close to .01. These values imply that all estimates are probably within ±.10 of the true probabilities, and most are within ±.08 if not ±.05. Thus, when using these probabilities to compare different score positions, distinctions between vowel transitions and rest transitions are very likely informative, as are distinctions between vowel transitions and consonant group transitions. Distinctions between a consonant group transition containing at least one unvoiced consonant and a consonant group transition with no unvoiced consonants probably are informative. However, distinctions among consonant group transitions with at least one unvoiced consonant are more likely the result of noise in the estimated probabilities, as are distinctions among consonant group transitions with no unvoiced consonants. Finally, note that the likelihood of confusing one note transition with another can be calculated from the estimated probabilities, just as for events based on pitch and spectral envelope, if the actual probabilities are assumed equal to the estimated probabilities.

Although the estimated probabilities appear reasonable for the specified conditioning variables, the appearance of probabilities around 0.5 clearly indicates the existence of omitted factors that affect observation of onsets. For instance, certain phonemes commonly are louder than others. Glides and liquids such as [w] and [l] are often much louder than voiced fricatives, affricates, and stops such as [v] and [g]. Conditioning probabilities on more precise phonetic classes might help. However, this approach would require examination of a much larger number of recorded performances. In general we expect at most one observed onset per note, not per phoneme. Besides depending upon phonemes, detection of onsets also depends upon the individual performer and the piece that is performed. For instance, some performers produce loud bursts when singing stops while others do not. Pieces with large dynamic ranges require a low amplitude threshold in soft sections, but this threshold may be exceeded by the amplitude of voiced consonants in the louder sections. Modeling onset distributions based on the dynamic range of the piece might help. Alternatively, an adjustable threshold might be applied based on dynamics for different sections of the piece or possibly some form of automated adaptation to signal levels. Finally, altering the score based on careful confirmation of the breaths and rests actually observed by an individual singer might raise the probability closer to 1 for the "Rest" transition class.

Score positions must be grouped into events based on observing note onsets. Several of the defined transition classes are based on the vowels, rests, and consonants that precede a note. These classes are defined to span the first 150 ms of the vowel for each note. The "End of Note" class spans the consonants at the end of notes that precede a rest. All other portions of a note are associated with the "Middle of Note" transition class. Obviously this simple partitioning is only an approximation to the actual occurrence of an onset at each point in the score. For the transition classes covering the start of notes, the chosen length of 150 ms is based on both the 100 ms of signal needed for pitch detection and use of the slowest possible tempo when converting scored pitches and phonemes into events. Since it is very unlikely that the performer will choose a tempo that is 1.5 times as fast as the slowest tempo, 100 ms of actual time following an onset will always correspond to less than 150 ms of score time. This approximation is not unreasonable given the broad definition of the transition classes.

In contrast, specifying that onsets are reported only within the vowel and not the preceding consonants is not entirely true. However, this simplification is a good approximation for the transition classes of "Vowel", "Rest", and "1 Consonant", and is only invalid by perhaps the duration of one consonant for the transition class "2 Consonants". Since these classes account for the majority of note transitions, the assumption that detection of an onset occurs only at the start of vowels will introduce limited errors in position estimation. Furthermore, a more precise method for specifying the transition classes of score positions would probably require categories that also account for the individual phonemes in the score.

The probabilities for observing note onsets fall into two extremes for most transition classes, indicating certain regions of the score are clearly distinguished from other regions based on observing or failing to observe a note onset. The relatively small regions of the score associated with note transitions further indicate that detection of an onset is likely to be informative. However, clearly detected note onsets offer limited assistance in discriminating the beginnings of different notes in the score. Nevertheless, they do offer information about score position that complements the previously considered observations of fundamental pitch and spectral envelope. In the next section, an explanation is provided for why combining observations of note onsets, fundamental pitch and spectral envelope will lead to improved tracking of vocal performances.

6.4 Improved Position Estimates Using Multiple Observation Types

In Chapters 4 through 6, three types of events for vocal scores are defined. The observation distributions estimated for these events cannot completely discriminate between all instances of the respective event type. Furthermore, the distributions do not include all relevant conditioning variables and complete accuracy of the estimates is questionable. Considering that arguments have been given for why each observation type in isolation cannot provide perfect discrimination of a vocalist's score position, it may not be obvious that combining multiple imperfect types of observations will lead to improved vocal performance tracking. In this section, motivation for expecting such improvement is provided, along with a quantitative method for comparing different sets of observation types and determining both the observations that should be expected to provide better score following and the extent of the improvement.

Recall that the stochastic score following model (in the continuous case) is defined by the following set of equations:

$$f_{I\,|\,D}\,(i|\;d = \mathrm{d}_1) = \int_{j=0}^{\|Score\|} f_{I-J\,|\,D}\,(i-j|\;d=\mathrm{d}_1)\cdot f_{Prior}(j)\;\partial j$$

$$f_{I\,|\,D,V}\,(i|\;d=\mathrm{d}_1,\;v=\mathrm{v}_1) = \frac{f_{V\,|\,I}\,(v=\mathrm{v}_1|\;i)\;\cdot\;f_{I\,|\,D}\,(i|\;d=\mathrm{d}_1)}{\int_{k=0}^{\|Score\|} f_{V\,|\,I}\,(v=\mathrm{v}_1|\;k)\;\cdot\;f_{I\,|\,D}\,(k|\;d=\mathrm{d}_1)\;\partial k}$$

where i = the performer's destination position
d = the estimated distance
v = the observation(s)
j = the performer's source position
f_{Prior} = the previous score position density function
$f_{I\text{-}J\,|D}$ = the distance density function
$f_{V|I}$ = the observation density function

The observation distribution, $f_{V|I}$, is multiplied by the result of convolving the stochastic description of the performer's previous position with the stochastic description of the amount of score performed since the position was last estimated. Ideally, the new stochastic estimate of the performer's position, $f_{I|D,V}$, should be a sharp distribution having most of its area near to the actual position of the performer. As previously mentioned, sharpness implies that the distribution has small variance over I and preferably is unimodal, where small is relative to both the possible values of I and the required accuracy of the position estimate. Since the result of the convolution equation always has an increased variance relative to the previous position density, the sharpness of the new score position density depends mainly upon the sharpness of the observation distribution, $f_{V|I}(v=\mathrm{v}_1|\;i)$. Thus ideally the observation distribution also should exhibit small variance over i and have most of its area near to the actual position of the performer.

In Chapter 4, a similar requirement was specified for insuring accurate estimation of the observation density. Specifically, $f_{V|I}(v|i=\mathrm{i}_0)$ should be unimodal and have small variance over the variable v. This requirement was justified by central limit theorems, which imply that the distribution will become essentially normal or lognormal if all primary factors (conditioning variables) are known. Suppose that the function $f_{V|I}$ is specified for a single observation type. As one reduces the variance of this function with respect to V, observations of the specified type may also provide more fine distinctions between score positions. In other words, the observation distribution may become sharper with respect to I as well. Specifying additional conditioning variables for the observation distribution and estimating the modified function may produce this effect. However, as previously and extensively discussed, the ability of a single observation to discriminate different positions has inherent limits. Furthermore, estimation of such extended functions may require larger samples, and the values for some important conditioning variables may be impossible to determine or estimate with high accuracy.

Consequently, one alternative is to incorporate more observation types, or conditioned variables, in the observation density. When adding conditioning variables, the central limit theorems can be used to argue that the observation density will sharpen over V if the added variables correspond to primary factors determining the observation. Since the new score position density, $f_{I|D,V}$, determines the score position conditioned on the observations, V, analogous arguments should also apply to this function. Now, since it is most intuitive to think of the observations as determined by the score position, it is less apparent that there necessarily exist arguments supporting a converse relationship. However, it is helpful to recall that the sharpness of the new score position density depends upon the sharpness of the observation distribution. As has been discussed, adding conditioned variables (observation types) corresponds to using a joint density for the observation density. This density can be rewritten as a product:

$$f_{V1,V2,V3|I} = f_{V1|I} \cdot f_{V2|I,V1} \cdot f_{V3|I,V1,V2}$$

If observation types are selected such that the product of their density functions is more likely to be sharper over I (perhaps even more normal or lognormal) than any individual function, then combining observation types will improve performance tracking. Observation types that effectively discriminate between different positions may produce this behavior. In this situation, adding observation types may improve position estimation accuracy beyond what is achieved by increasing accuracy of the density functions for individual observation types. In some instances, including more types of observations and estimating new density functions actually may be easier than including new conditioning variables and re-estimating the density functions previously defined. For instance, estimating density functions for new observations may entail additional processing of previously collected recordings, but may not require additional recordings in order to achieve reasonable accuracy.

To help the intuition, Figures 6-2 through 6-4 depict several examples of the sharpening and "flattening out" of $f_{V|I}$ when incorporating multiple observations. These figures are based upon the actual distributions estimated for the three types of observations already discussed. The portion of score considered is an excerpt from "Happy Birthday". Consider the case when the performer has begun to sing the vowel in the third syllable, "Birth-". Figure 6-2 shows observation density functions over score position when observing a fundamental pitch of E, a spectral envelope (vector codebook entry) commonly

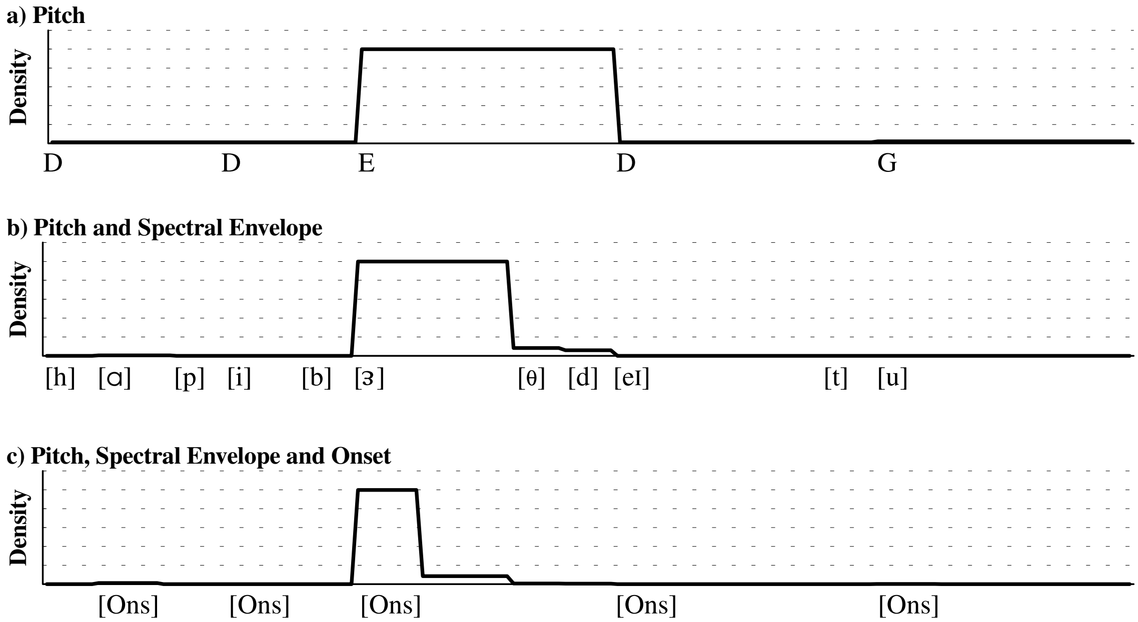


Figure 6-2. The observation density function, $f(v|\ i)$, based on observing "E", [ɜ], and an onset, given part of the score for "Happy Birthday". Graph a) shows the distribution based on observing a fundamental pitch of "E". Graph b) shows the distribution based on observing both a fundamental pitch of "E" and a spectral envelope likely to be seen when the vowel [ɜ] is sung. Graph c) shows the final distribution when observing fundamental pitch and spectral envelope as for graph b) as well as a note onset.

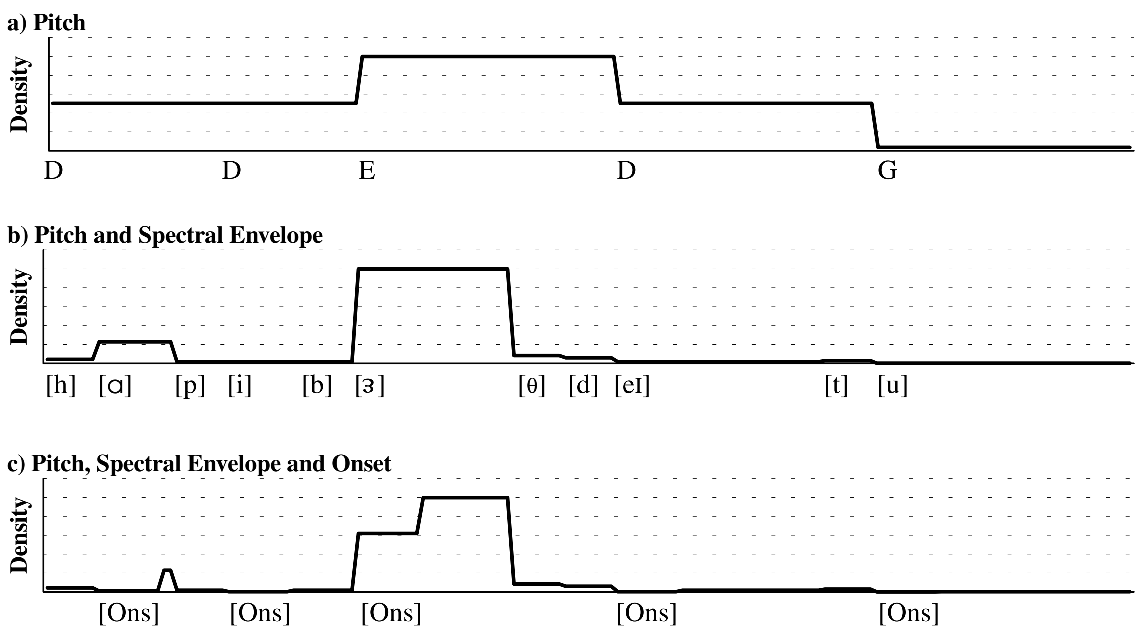


Figure 6-3. The observation density function, $f(v|\ i)$, based on observing "E^b", [ɜ], and no onset, given part of the score for "Happy Birthday". Graph a) shows the distribution based on observing a fundamental pitch of "E^b". Graph b) shows the distribution based on observing both a fundamental pitch of "E^b" and a spectral envelope likely to be seen when the vowel [ɜ] is sung. Graph c) shows the final distribution when observing fundamental pitch and spectral envelope as for graph b) as well as lack of a detected note onset.

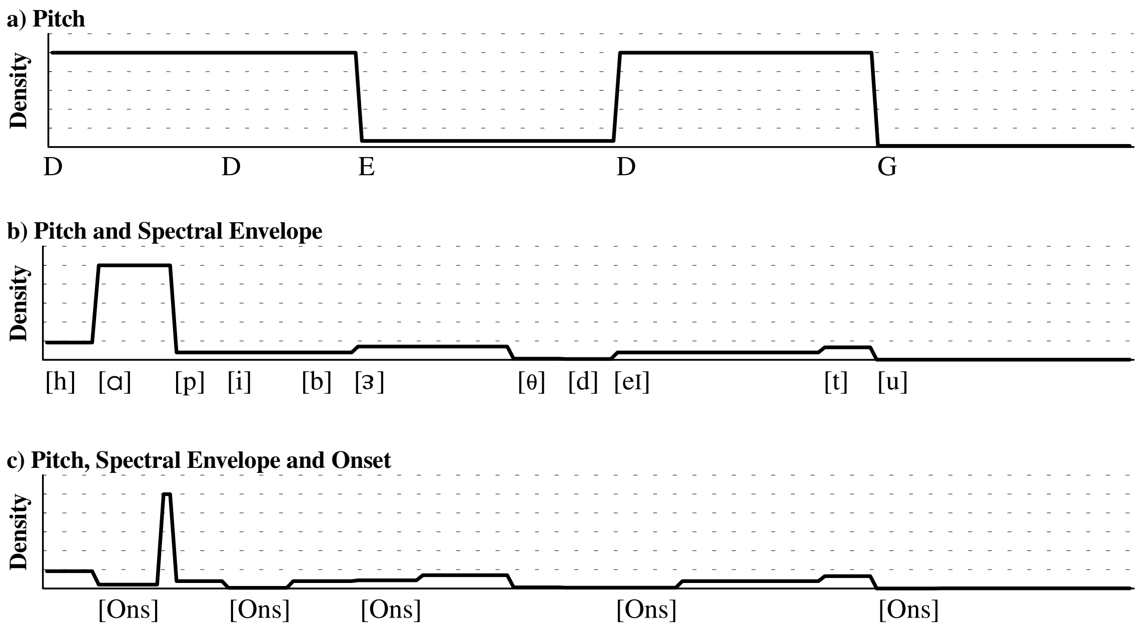


Figure 6-4. The observation density function, $f(v|\, i)$, based on observing a "D", [ɜ], and no onset, given part of the score for "Happy Birthday". Graph a) shows the distribution based on observing a fundamental pitch of "D". Graph b) shows the distribution based on observing both a fundamental pitch of "D" and a spectral envelope likely to be seen when the vowel [ɜ] is sung. Graph c) shows the final distribution when observing fundamental pitch and spectral envelope as for graph b) as well as lack of a note onset.

observed for the vowel [ɜ], and an onset. Note that the density based on all observations is the sharpest in the correct region. Figure 6-3 shows observation density functions over score position when observing the same spectral envelope, but a fundamental pitch of E^b and no note onset. These observations are not unexpected for the start of the third syllable, but are more likely to be observed around the middle of the third syllable. Note that the distributions that combine observations indicate this more clearly than the distribution based solely on pitch. Figure 6-4 shows observation density functions over score position when observing the same spectral envelope and no note onset, but a fundamental pitch of D. This combination of observations would be less common when a singer is starting the third syllable. While no depicted density function clearly distinguishes the correct region, note that the density function combining all observations contains a more even distribution of area than the other two density functions. While this does not determine the correct score position, it does increase the likelihood that the result of the convolution will determine the position estimate. If the convolution estimates are generally reasonable, this behavior is appropriate when observations cannot distinguish the correct position.

While a few diagrammed examples assist in understanding how combining observations can affect the score position density, they do not provide conclusive or even convincing evidence that score position estimation will improve on average when using several specific observation types. Such proof requires defining necessary and sufficient conditions for the product of the observation density functions to be sharper around the performer's actual position than are the individual functions. Obviously, if all observation types always assign highest density to the actual position, the product will become sharper. However, if some observation types sometimes assign highest density to another position, then the average behavior must be considered.

For a very restricted case, the expected behavior can be examined analytically. Consider a set of observation types whose respective observation density functions always have the same maximum value, although the point of maximum density may differ for each function. Also assume that every other point within each density function is assigned equivalent density. For each of these observation types, consider the likelihood that the observation density assigns highest density to the wrong position. Furthermore, consider that these likelihoods are identical and independent. If all of these conditions are true, then for a single joint observation, the probability that a fixed number of the observation types will fail to assign highest density to the performer's actual score position is defined by the following binomial distribution:

$$\mathrm{P}\,[X = x] \;=\; \binom{n}{x} p^{x}(1-p)^{n-x}$$

where n specifies the number of observation types and p specifies the probability of assigning highest density to the wrong position.

Now in the worst case, each observation type that assigns maximum density to a wrong position will assign this maximum to the same position. If half or more of the observation types do this, then the combined observations will fail to distinguish the correct position. Given a single joint observation, the probability that half or more of these observation types will assign highest density to the wrong position is defined by the following summation:

$$\mathrm{P}\,[X \geq \tfrac{n}{2}] \;=\; \sum_{k=\lceil \frac{n}{2} \rceil}^{n} \binom{n}{k} p^{k}(1-p)^{n-k} \qquad \textbf{[6–1]}$$

Now also assume that for a single joint observation, each observation type that assigns highest density to the wrong position always assigns the maximum density to the exact same position. In this situation, the given summation defines the percent of instances when the joint density function for all observation types assigns highest density to the wrong score position. Thus for the considered situation, equation 6-1 defines the likelihood of confusing another score position with the performer's actual score position based on a single joint observation.

Table 6-2 contains values of this summation for various values of p and n. While the assumptions underpinning this calculation are unlikely to apply in real statistical modeling problems, the values of the summation demonstrate several important facts. First, for fixed n, the probability of assigning highest density to the wrong position of course decreases as p decreases. Thus, improving the position estimation accuracy of the individual observations without adding new observations will improve overall position estimation. Second, improvements can be obtained when increasing the number of observation types for a fixed p. However, the probability does not decrease monotonically with n. Sometimes addition of at least 2 observation types is required to obtain an expected improvement in position estimation. This effect results from using the ceiling of $\frac{n}{2}$ in equation 6-1. Nevertheless, improvement in estimation does occur as the number of observation types increases significantly. Furthermore, the degree of expected improvement when adding one or two observation types is sometimes superior to or comparable to achieving a .05 reduction in the likelihood that each observation type will assign highest density to the wrong position.

Even for this extremely simple case, the improvement in position estimation is a nontrivial function of both the number of observation types and the likelihood that each observation type will not distinguish the performer's actual position. Thus, selection and modeling of observation types must consider the tradeoff between increasing the number of observation types and using only observation types that are more likely to distinguish the actual position. Adding observations that are poorly modeled and offer little additional position discrimination may actually reduce overall estimation accuracy. Even worse, for many distribution functions, the modeling of distributions including conditioning variables that are uncorrelated with the conditioned variables generally is most difficult and requires a larger sample to

Table 6-2. Results of calculating equation 6-1 for various values of n, the number of observation types, and p, the probability of assigning maximum density to the wrong score position.

n	$p = .01$	$p = .05$	$p = .10$	$p = .15$	$p = .20$	$p = .25$
2	0.0199	0.0975	0.1900	0.2775	0.3600	0.4375
3	0.0003	0.0072	0.0280	0.0608	0.1040	0.1562
4	0.0006	0.0140	0.0523	0.1095	0.1808	0.2617
5	0.0000	0.0011	0.0085	0.0267	0.0579	0.1035
6	0.0000	0.0022	0.0159	0.0474	0.0989	0.1694
7	0.0000	0.0002	0.0028	0.0122	0.0334	0.0706
8	0.0000	0.0004	0.0050	0.0213	0.0563	0.1138
9	0.0000	0.0000	0.0009	0.0056	0.0196	0.0489
10	0.0000	0.0000	0.0016	0.0098	0.0328	0.0781

achieve high accuracy. Recall that for a Bernoulli distribution, accurate estimation of $p = .5$ is the worst case, and accurate estimation of a uniform distribution using a histogram also requires a large sample.

The simple case just examined makes several unrealistic assumptions. For instance, it is unlikely that all defined observation types will have identical likelihood of assigning highest density to the wrong position. Certainly the observation types developed for vocal performance tracking do not satisfy this assumption. Relaxing these assumptions will of course lead to a more complicated calculation for determining expected improvements when combining observation types. However, such a calculation is possible providing that the density functions are known for all considered observation types. It is based upon the likelihood of confusing two points in the score, as was considered for each of the observation types previously defined:

$$\kappa(i,j,f_{V|I}) = \int_{v=v_0}^{v_n} f_{V|I}(v|i) \cdot [f_{V|I}(v|j) \geq f_{V|I}(v|i)]\, \partial v$$

The function κ defines the likelihood of assigning equal or higher density to point j than to point i for a single observation, v, having the given observation density function. If the observation density, $f_{V|I}$, is a joint density function over several observation types, the calculation is adjusted to compute multiple integrals, one integral per observation type:

$$\kappa\left(i,j,f_{V1,V2|I}\right) = \int_{v1=v1_0}^{v1_{n1}} \int_{v2=v2_0}^{v2_{n2}} f_{V1,V2|I}(v1,v2|i) \cdot \left[f_{V1,V2|I}(v1,v2|j) \geq f_{V1,V2|I}(v1,v2|i)\right] \partial v2\, \partial v1$$

This equation corresponds to equation 6-1 in a more general case. For a given score, the expected likelihood that one point will be confused with another is then given by:

$$E[\kappa] = \frac{1}{\|Score\|^2} \int_{i=0}^{\|Score\|} \int_{j=0}^{\|Score\|} \kappa(i,j,f_{V|I}) \cdot [j \neq i]\, \partial j\, \partial i \qquad \text{[6-2]}$$

Note that this calculation assumes equal likelihood that the singer is at any score position when an observation is reported. Finally, if κ is substituted and the order of integration changed, equation 6-2 can be rewritten as follows:

$$E[\kappa] = \frac{1}{\|Score\|^2} \int_{i=0}^{\|Score\|} \int_{v=v_0}^{v_n} f_{V|I}(v|i) \int_{j=0}^{\|Score\|} [j \neq i] \cdot [f_{V|I}(v|j) \geq f_{V|I}(v|i)]\, \partial j\, \partial v\, \partial i$$

which defines the expected percent of the score that will be confused with (have equal or higher density than) the performer's actual score position.

The expected likelihood of confusing one point with another in a specific score can be used to approximate the expected likelihood of confusing one point with another over all possible scores. This latter value should be expected to correlate with the ability of one or more observation types to discriminate score position. One way to approximate this value is to obtain statistics for the distribution of conditioning variables (*e.g.*, phonemes, transition classes) that are substituted for i in the observation distributions. Such statistics would provide a meaningful measure over i and over j relative to the observation distribution. This measure could be used in a modified form of equation 6-2 to approximate expected likelihood of confusing two points over all possible scores. Alternatively, equation 6-2 could be applied to a sample of scores and the average over the score specific values computed. It is likely that such a sample of scores would be needed to estimate the statistics for the first approach anyway, unless it is reasonable to assume uniform distribution of the conditioning variables.

The expected likelihood of confusion defined over a set of scores can be viewed as a measure of the sharpness and the distribution of area for the observation density, $f_{V|I}$, in actual circumstances. It is based on the distributions of both the conditioning variables substituted for I and actual observations. The calculated value should correlate well with both the average sharpness of the observation density and any metrics for assessing tracking accuracy. One such metric is defined in the next chapter. The degree of correlation depends upon the importance of observations in discriminating score position (compared to rate and elapsed time) and whether or not the model and estimated density functions adequately characterize the real world.

Adequate characterization of the real world again relies upon accurately estimating observation densities and including all primary factors in the density functions. Fortunately, this requirement can be satisfied by applying good statistical estimation techniques and carefully considering all model assumptions. In addition, including all primary factors may not be as important as accuracy if the calculation is used to estimate average behavior. As long as distributions are accurately estimated over the selected factors and a sufficiently large number of scores is examined, the resulting expected likelihood calculation will be correlated with the averages of metrics applied to actual performances. However, the calculation may be less correlated with metrics applied to individual scores or performances. Also, when comparing calculated values, larger differences between the values will offer stronger evidence that the corresponding metrics also will differ.

The defined calculation does not consider the effect of the convolution result on the final position estimate. Thus, it does not account for errors resulting from poor tempo estimation. Consequently, the calculated value will not be correlated completely with the sharpness of the new position density and any proposed metrics. However, such a limitation may not affect comparisons based on the calculation. Adding observation types to the score-following model will not alter the convolution, so the convolution's effects on the sharpness of the score position density may be similar, regardless of the observation types. Thus, once again the average of metrics applied to a sufficiently large and unbiased set of performances and scores will remain correlated with the calculation.

To the extent that the expected likelihood of confusion is correlated with the actual tracking accuracy of a statistical model, the calculation will have several beneficial uses. First, it will provide a way to compare different versions of the statistical model prior to actual testing. It may often be more expedient overall to use the calculation as a first pass evaluation rather than proceeding immediately to actual testing, possibly iterating through several estimated models. Although for vocal performance tracking it is feasible to run tests with recordings, comparing estimated positions with actual positions is difficult to define and time consuming to execute. Also, collection of sufficient data for model estimation is clearly a problem. It would be preferable to use all available recordings for model estimation, test with live performances only, and limit the amount of testing.

Thus and second, the calculation also can be used to determine when testing is worthwhile. Even when the tracking model is tested with recorded performances, it is necessary to do at least some testing with live performances anyway. Since testing of live performances is a time consuming and drawn out process, one must limit the extent of this testing and should be as certain as possible of success prior to performing the trials. The calculation could be used to decide whether expected improvements due to an extended or revised model are worth the trouble of live performance testing. As already mentioned, this evaluation could be done in place of testing with recorded performances.

Third, if the calculation and test results are correlated, the calculation can add support for believing the test results. Live performances are not repeatable. It is useful to have additional support for believing that observed differences in tracking accuracy result from differences between the compared models and not from differences in the live performances. Also, in general only a handful of live performances can be tested in a reasonable period of time, since only a limited number of performers will be available to participate in the tests.

Finally, the results of the calculation may provide a more accurate assessment of the average tracking accuracy of a model. It is easier to generate large numbers of scores than to obtain large numbers of performances. Providing that the observation densities are reasonably accurate with respect to the conditioning variables, the value calculated from a large number of scores may be preferable to the results obtained from only a handful of performances. Thus, although testing is certainly still necessary to confirm or refute utility of a score-following model, an analytical calculation that estimates likelihood of confusing score positions can help to confirm the relative success or limitations of any examined versions of the score-following model.

In summary, this section has considered how to determine whether or not incorporating multiple types of observations into the score-following model will improve overall performance tracking. The abstract answer given is that it depends upon the sharpness and distribution of area exhibited by the estimated score position density function. A sharpening of this function is equivalent to a sharpening of the observation density function, $f_{V|I}$, over the variable I. Such a sharpening can occur through the addition of observation types. However, this sharpening can depend upon both the number of additional observation types and the ability of the new observation types to discriminate the performer's actual score position. While there has been no formal specification of either necessary or sufficient conditions for sharpening of the observation density, an analytical technique for assessing the expected likelihood of confusing two positions has been described. This technique relies on reasonable estimation of the individual observation densities of interest and either a set of scores or comprehensive statistics for the conditioning variables in the distributions. It is expected that the results of this computation will be correlated strongly, in general, with sharpness of the score position density and metrics that assess accuracy of score position estimation. In the next chapter, this expectation will be fulfilled for the case of score following using the stochastic models already described.

Now, a comment can be made about expecting improved score position estimation on average when combining all three of the previously discussed observation types. First, note that each of the selected observation types offers some ability to distinguish score position. The estimated density functions based on the respective conditioning variables are not uniform. In some instances, the distributions over the observation values are fairly sharp. Second, the observation types considered offer some complementary ability to distinguish positions. Specifically, each observation type has the potential to distinguish different regions of the score and the regions differ amongst the observation types. Third, it can be expected that the likelihood that each observation type will assign highest density to the wrong score position is not correlated too strongly. The occurrence of a less likely value for one observation type, given the performer's actual score position, does not imply necessarily or even probably that the observation values for the remaining observation types will be unlikely values. Finally, while the distributions estimated for the observation types do not include all pertinent conditioning variables, overall they have been estimated from reasonable size samples and probably include only a limited number of extreme inaccuracies relative to the chosen conditioning variables. Considering these facts in light of the preceding discussion, it is reasonable to expect that average tracking accuracy will improve when using all three types of observations. As presented in the next chapter, both the calculations for expected likelihood of confusion and the metrics applied to actual performances support this conclusion.

While not attempted as part of this work, it may be possible to provide a more formal but useful definition of necessary or sufficient conditions for increased sharpness, as well as methods for determining the rate of increase. Factorization of the joint observation density may be the key. For instance, it is easily shown that a product of several Gaussian functions having identical mean and variance is also a Gaussian with smaller variance. Normalizing a Gaussian of course yields a normal density. Similarly, many central limit theorems discuss convergence of the distribution of a sum or product of independent random variables. As previously discussed, the distribution of a sum of independent variables is the convolution of the individual distributions. Convolution of distributions corresponds to multiplication of their Fourier transforms, or characteristic functions. While the convolution generally produces a function having greater variance than either of the individual functions convolved, the corresponding product of the Fourier transforms (*i.e.*, the transform of the convolution result) generally will have smaller variance. Proofs of central limit theorems often are accomplished by showing convergence of the product of the transforms. For instance, since the transform of a Gaussian is also a Gaussian, the convergence of repeated convolution of distributions to a wide-variance normal curve often is proved by demonstrating the convergence of the product of the transforms to a small-variance normal curve. Also, it is often shown that increasing the number of distributions improves this convergence. Although these proofs usually rely on several assumptions not valid for real observation densities, such as that the individual densities are identical, proofs based on reasonable assumptions may be possible. Such results might lead to definitions of sufficient and necessary conditions for increased sharpness that could be applied in initial identification and selection of potential observation types.

The work described in this document has considered only three observation types for use within the stochastic score-following model. If combining three observation types improves performance tracking, it might be useful to find ways of rapidly but effectively increasing the number of observation types. However, each of the investigated observation types has required different signal processing techniques, data analysis, and different density functions to be defined and estimated. Specifying the joint observation distribution has required considerable effort. Thus, having to produce many distinct, independent observation types based on completely different conditioning variables may be limiting. Now one intuitively silly alternative would be to report multiple instances of the exact same observation type, leading to a joint density such as the following:

$$f_{V|I} = f_{V1,V1|I} = f_{V1|I} \cdot f_{V1|I,V1}$$

If the product on the right actually produced a sharper distribution, this approach would be statistical modeling's equivalent of perpetual motion. Fortunately, however, $f_{V1|I,V1} = 1$ always (or more properly, equals a unit impulse centered at $v1$, $\delta(V1-v1)$), and any accurate estimation process of course will produce this equivalence. Thus, the joint distribution will reduce to $f_{V1|I}$ and no advantage will be gained.

An only slightly less silly approach would be to try using previous observations in addition to any new observation. Now in the score-following model, it is assumed that the recent observation is dependent upon the destination position, *i*, so the immediately preceding observation is dependent upon the source position, *j*. Thus, distributions involving previous observations could not be extricated from the convolution step without making very questionable assumptions. In addition, since the distribution over the source position, $f_{Source}(j)$, technically includes previous observations as conditioning variables, using distributions over previous observations without making unreasonable assumptions would lead to situations similar to the one just demonstrated, where the additional density functions always equate to 1 (a unit impulse). Now the only reason this option is slightly less silly than the previous suggestion is that if we did properly redefine our score-following model to include the previous observation and condition it on both *i* and *j*, then we actually might obtain better position estimates. However, this process is likely to be complicated and may not yield a model that can be estimated from available data and also computed in real time.

Now one viable alternative might be to use slightly altered versions of the same sensors. For instance, the distribution estimated for fundamental pitch showed a noticeable spread relative to the scored pitch. One factor contributing to the spread is the singer's vibrato, especially since vibrato periods often exceed the 100 ms of signal processed for each reported pitch. Simultaneously considering observations from a duplicate pitch detector that applies median smoothing over an increased amount of signal might serve to counteract vibrato effects. Puckette (1995) has reported that use of both short term and long term average pitch is helpful. Now probably it would not be safe to assume independence between the two pitch observations, and the density function for one pitch estimate would have to be conditioned on the other pitch estimate. To the extent that the two values are strongly correlated, the additional observation will not offer much ability to discriminate positions. If the density function is defined and estimated accurately, this situation will be made apparent by the density function's shape. However, if the long term median pitch does differ in some cases from the short term median pitch, the resulting joint distribution might provide slightly improved position discrimination. Even though long term medians may not always be available, using both pitch estimates when possible might be better than using only one or the other. In addition to long term and short term medians, building multiple pitch detectors based on different analysis techniques could be helpful. For instance, applying harmonic matching to the short time spectrum might produce estimates that compensate for some errors introduced by the filter bank approach. Although such additional observations may offer limited improvements, they may at least marginally improve position estimation without requiring additional data collection. The previously described analytical calculation could be used to compare numerous possibilities prior to actual testing.

The next chapter contains a description of tests run on various versions of the stochastic score-following model. These tests used the score follower as part of a complete automated accompaniment system. They consider the accuracy of the tracking and the accompaniment synchronization for both recorded and live performances when using different combinations of observation types. The results demonstrate that combining observations does improve tracking both for individual performances and on average. The formal estimate of a tracking system's expected performance, defined in this chapter, is calculated for each score and compared against the empirical results. This comparison shows good correlation between the calculations and the actual results, both among individual performances when using the same observation types and across averages over performances when using different sets of observation types. Finally, the discussion identifies situations where serious errors in position estimation occur and relates them to the model assumptions and estimation techniques. These results are used to identify the most important work for subsequently enhancing automated tracking of vocal performances.

Chapter 7

Application and Evaluation of the Stochastic Score-following Model

7.1 Overview

This investigation of vocal performance tracking for automated accompaniment has proceeded in several stages. First, musical accompaniment and vocal performances were considered, providing motivation for a stochastic method of tracking vocal performers. Next, a stochastic model for performance tracking was developed along with an approximate but efficient implementation. The defined implementation required the specification and estimation of several component density functions. These functions included an observation density that, due to the requirements for robust vocal performance tracking, had to incorporate multiple observation types. During the estimation process, much attention was given to the accuracy of the estimated density functions. Finally, it was argued that combining observation types would lead to improved performance tracking. A calculation was defined over the estimated observation densities and example scores, in order to estimate relative improvements in tracking accuracy that are obtained by combining observation types.

In this section, an implemented accompaniment system is described. This system incorporates the stochastic score-following model. Details of implementing this system are provided, including the method for controlling real-time playback of the accompaniment based upon score position and tempo estimation. Metrics are defined for assessing both the tracking accuracy and the synchronization between the accompaniment and live performer. These metrics have been applied to determine average accuracy during individual performances with the accompaniment system, as well as to determine average accuracy across multiple performances. Consideration is given to changes in accuracy that occur when the tracking system uses different observation types, and also when the accompaniment control relies on different point estimates of score position. The metrics have been applied to both recorded and live performances. The results demonstrate the utility of the stochastic tracking technique, confirm improved tracking accuracy when using multiple observations, and identify the most significant causes of both tracking and synchronization errors. Finally, results of actual performance trials are compared with the expected results based on calculating the expected likelihood of confusion previously defined. This comparison demonstrates the validity of using this computation for a quick assessment and comparison of different versions of the stochastic score-following model.

7.2 The Automated Accompaniment System

Figure 7-1 contains a diagram of signal and data flow through all hardware and software components in the automated accompaniment system. Signal from a live performance or recording passes through an analog mixer that boosts frequencies around and below 220 Hz, sharply attenuates frequencies below 75 Hz, and alters the signal level. The signal is patched from the mixer into the sound card on a personal computer. It is then lowpass filtered and sampled at a rate of 16 KHz. The operating system interacts with the sound card and provides buffers of digitized signal to the signal processing software. These buffers undergo blocking and thresholding to separate pitched from unpitched signal. The detectors for the three observation types process the blocks of signal, and the combined results are time stamped and marshaled into a MIDI (Musical Instrument Digital Interface) message (MIDI Manufacturers Association 1989). The observations are sent to a second computer via a direct MIDI connection between the sound cards.

Once received by the second computer, the observation message is passed from the sound card, through the operating system, to the CMU MIDI Toolkit. This toolkit receives and decodes MIDI messages, providing a framework for interactive music applications (Dannenberg 1986; Dannenberg 1993a; Dannenberg 1993b). The observation message is passed to software that implements score following via the stochastic model, estimation of the performer's score position and tempo, and control of the accompaniment. Position and tempo changes are passed to the MIDI Toolkit. This toolkit includes routines for handling real-time playback of a musical score via MIDI. These routines allow real-time adjustment of the current score position and tempo. In addition, since the musical score is represented internally as a sequence of function calls, the toolkit permits user definition of arbitrary functions and inclusion of arbitrary function calls as part of the musical score. The MIDI events in the score are sent via a direct connection from the second computer's sound card to an external synthesizer. This device generates actual audio based on the MIDI events. The audio output from the synthesizer is patched into the analog mixer where levels can be adjusted prior to patching the sound to speakers and recording devices.

The signal processing software resides entirely on the first computer. This machine contains a 66 MHz Intel Pentium processor. Recall that the sound signal is processed in 33 ms blocks and synchronized

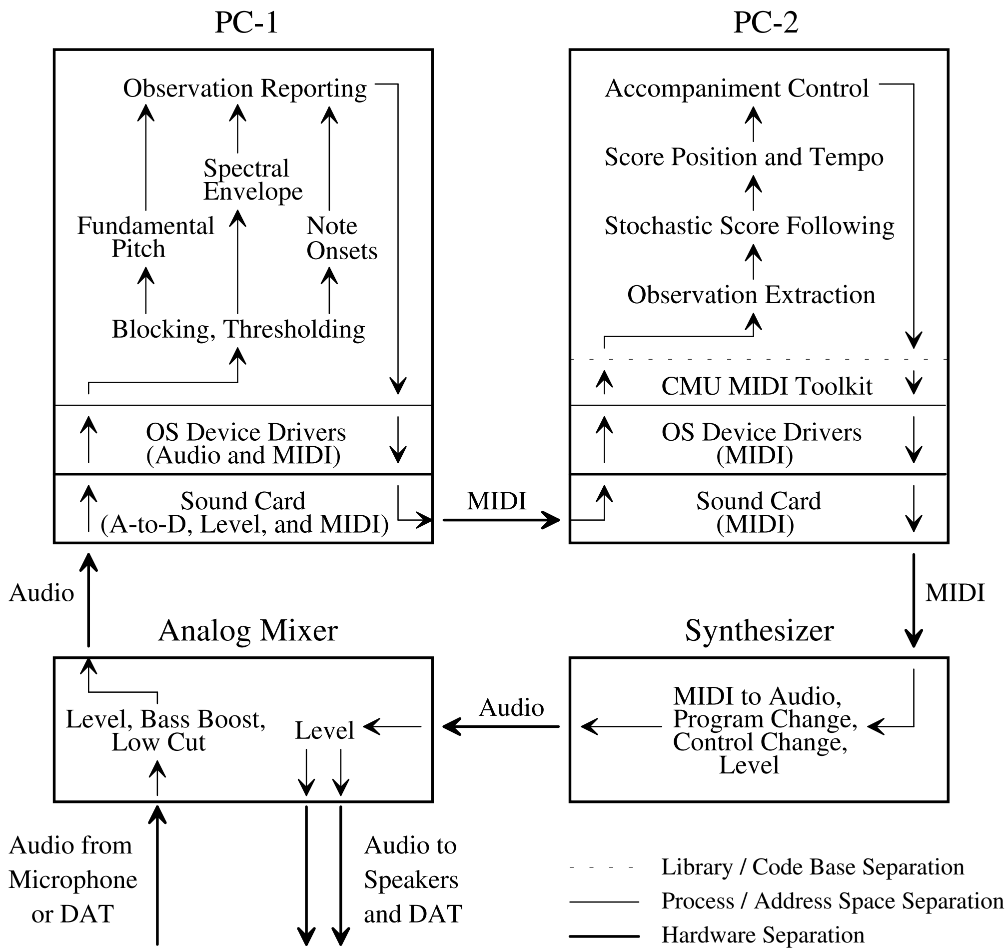


Figure 7-1. Diagram of signal and data flow in the automated accompaniment system.

reports for all three observation types occur every 100 ms during a sustained tone. Under these conditions, elapsed time for all signal processing was measured at less than 30 ms.

Position estimation and accompaniment control are implemented on the second computer. Figure 7-2 presents the equations for the final score-following model that includes all three observation types. This model is implemented using the method based on the fast Fourier transform, as described in Chapter 2. All scores were sampled using a 10 ms interval along the score position dimension. In Chapter 3, this interval was determined as sufficient to maintain acceptable errors when numerically computing the convolution. A window size of 512 samples is used, corresponding to just over 5 seconds of idealized score time. Consequently, the convolution is calculated using a Fourier transform with 1024 samples. The second computer contains a 166 MHz Intel Pentium processor. The implemented score-following model using a 1024-point transform requires 18 ms of computation time. This includes the real-time

Vocal Performance Tracking Model:

$$f_I^{t1}(i) = f_{I \mid J,\mathrm{D},V}(i \mid j = j_0,\, d = d_0,\, v = v_0[Pitch, Envelope, Onset])$$

$$\cong \frac{f_{V \mid I}(v = v_0[Pitch, Envelope, Onset] \mid i) \cdot C(i,j,d = d_0)}{\int_{i=0}^{\|Score\|} f_{V \mid I}(v = v_0[Pitch, Envelope, Onset] \mid i) \cdot C(i,j,d = d_0)\, \partial i}$$

where $C(i,j,d = d_0) \cong \int_{j=0}^{\|Score\|} f_{I-J \mid D}(i-j \mid d = d_0) \cdot f_I^{t0}(j)\, \partial j$

j = source position of the performer.

i = destination position of the performer.

d = estimated score distance traversed by the performer.

v = newly reported observation.

Distribution of Actual Score Performed:

$$f_{I-J \mid D}(i-j \mid d = d_0) \cong f_{I-J \mid R,\Delta T}(i-j \mid r = r_0,\, \Delta t = \Delta t_0) \cong \frac{1}{(i-j)\sigma\sqrt{2\pi}}\, e^{\frac{-(\ln(i-j)-\mu)^2}{2\sigma^2}}$$

where $\mu = \ln r_0 - \frac{1}{2}\sigma^2 + \ln \Delta t_0$

$\sigma^2 = \ln\left(\frac{1}{.02948\Delta t_0} + 1\right)$

$\Delta t_0 = t1 - t0$

r_0 = estimated average rate over preceeding ~3 seconds.

Distribution of Observations:

$$f_{V \mid \mathrm{I}}(v = v_0[Pitch, Envelope, Onset] \mid i) \cong f_{V \mid \mathrm{ScoredPitch}}(v = v_0[Pitch] \mid ScoredPitch(i)) \cdot$$

$$f_{V \mid \mathrm{ScoredPhoneme}}(v = v_0[Envelope] \mid ScoredPhoneme(i)) \cdot$$

$$f_{V \mid \mathrm{TransitionClass,\ VoicedConsonants}}(v = v_0[Onset] \mid TransitionClass(i), VoiceConsonants(i))$$

Distribution of Observed Fundamental Pitch:

$f_{V \mid \mathrm{ScoredPitch}}(v = v_0[Pitch] \mid ScoredPitch(i)) \cong$ probabilities from Table 4-1

Distribution of Observed Spectral Envelope:

$f_{V \mid \mathrm{ScoredPhoneme}}(v = v_0[Envelope] \mid ScoredPhoneme(i)) \cong$ distributions in Figures 5-6 and 5-7

Distribution of Observed Note Onsets:

$f_{V \mid \mathrm{TransitionClass,\ VoicedConsonants}}(v = v_0[Onset] \mid TransitionClass(i), VoicedConsonants(i))$

$\cong$ probabilities from Table 6-1

Figure 7-2. Stochastic score-following model using fundamental pitch, spectral envelope, note onsets, rate, elapsed time, and source position.

generation of samples for the lognormal distance density, multiplication of the observation distributions, and normalization of the result.

Real-time processing proceeds in the following manner. On the first computer, buffers of audio signal are filled by the sound card driver and passed to the signal processing software through the operating system. The buffer size is set to encompass 33 ms of signal, the amount of signal processed as a single block by the detectors. Once received by the signal processing software, each block is immediately time-stamped and processed. For every third consecutive block containing signal of sufficient amplitude (according to the thresholding technique previously described), values for fundamental pitch, spectral envelope, and note onset are generated and marshaled into a MIDI message, along with the time stamp of the last processed audio buffer. This message is immediately sent to the tracking software running on the second computer.

Once received by the second computer, the message is decoded and the information is provided to the stochastic tracking software. A score position density function is computed. Using this function, a point position estimate is then generated. This estimate in combination with previous estimates are used to estimate tempo. Finally, the position and tempo are passed to the accompaniment control system which compares this information to the current accompaniment control settings. Settings are altered as appropriate to adjust the playback of the accompaniment. Accompaniment performance occurs continuously. The incoming MIDI messages effectively initiate an interrupt of the performance, causing the MIDI toolkit to suspend playback until the message is completely processed by the score following and accompaniment control software. However, such interruptions do not affect the toolkit's assessment of current time, which is based on the internal computer clock. Since the entire position estimation and performance adjustment process requires around 20 ms, hesitations in playback will be at most this large and will occur only if an observation message is received immediately before the next note is to be performed.

The dominant contributors to accompaniment system latency are the signal processing and the stochastic score following. Score position estimation and accompaniment control require negligible time, as does message passing via the direct MIDI connection. Since the sample buffer size is set to the exact block size processed, and since no other applications are active, we assume negligible delay in receiving each filled sample buffer via callback from the operating system. Consequently, total latency for the accompaniment system can be estimated by the sum of required processing times for signal processing and score following, around 50 ms. For accompaniment control, this delay is not unreasonable. In addition, the accompaniment control system extrapolates the position estimate to account for the latency, using a linear model incorporating the estimated tempo:

$$RevisedPosEstimate = PosEstimate + Tempo \times Latency$$

This approach limits the effects of the latency on the accompaniment control. Also, the magnitude of the error introduced into the position estimate is well below the overall accuracy of the tracking system.

Since the implemented score-following model uses a window of limited duration and the signal processing does not report observations during silence, the tracking system must deal explicitly with long rests. The necessity of dealing with long rests was first mentioned in Chapter 3. As described, the approach to handling long rests relies on accurate identification of extended silences. If the elapsed time between reported observations exceeds 2 seconds, the score-following system recognizes an extended rest. This period of time is sufficiently long to exclude the possibility that intervening consonants, catch breaths, or delays due to system latency will be mistaken for a rest. Also, it is short enough that observations spaced by less than this amount will not violate the estimated convolution model. When a rest is recognized, the source position density function is altered prior to application of the model. Specifically, the density is changed to a unit impulse centered at the end of the first rest following the last estimated position of the performer. The density is modified in this manner once a new observation is reported, prior to estimating position. Calculation of the destination position density function then proceeds according to the model, but using the revised source position density and an elapsed time of 100 ms (the amount of signal processed per observation). The initial source position density and elapsed time also are specified in this manner at the beginning of the performance.

The accompaniment control system requires a point estimate of the current score position. To provide such a position estimate, the tracking system identifies the 100 ms region of the score encompassing the most probability. This region corresponds to the portion of score where the area under the score position density function is maximized. The center of this region is taken as a point estimate of the performer's current position. Many studies in music perception have demonstrated that generally, in performances of melodies, timing deviations below 50 ms are difficult to perceive (Clarke 1989; Schulze 1989). While these experiments consider only melodies, it is reasonable to assume that perceivable timing deviations between two performers are of similar magnitude. If the score position density is interpreted as defining the probability distribution over the singer's actual position given all processed information, then synchronizing the accompaniment to the 100 ms region of the score position density encompassing the most area can be viewed as maximizing the number of times that the performer and any listeners will not perceive a timing error. In the experiments subsequently described, accompaniment control is always based on this position estimate. However, two other point position estimates are calculated and assessed for tracking accuracy. These estimates include the mean of the score position density function (the mean squared error estimate) and the mode of the density.

Both the score-following model and the accompaniment control system require an estimate of the performer's current tempo. Tempo estimation proceeds according to the method described in Section 3.5. To recap, a buffer of recent position estimates is maintained along with time stamps indicating when the estimates were produced. This buffer contains at most one estimate for each note in the score—the most recent estimate indicating a transition from the preceding note to the given note. Tempo estimates are produced using position estimates that span a specified minimum time. In general, this minimum elapsed time is around 3 seconds (specific minimums for each set of experiments are provided in subsequent sections). The performer's current tempo is estimated using the following calculation:

$$Tempo = \frac{Position_{Time2} - Position_{Time1}}{Time2 - Time1}$$

New tempo estimates are generated only when position estimates indicate a transition to a new note. The buffer of position estimates is emptied whenever a new position estimate exceeds the previous position estimate by both two notes and a distance indicating that the performer has increased tempo by at least a factor of two. The latter situation can occur if the performer jumps ahead in the score or the position estimates already in the buffer are incorrect. If a new position estimate precedes (by score position) the estimates already in the buffer, the buffer is rolled back to the first estimate that precedes the new estimate by at least one full note. The new position estimate is then added to the buffer and a new tempo estimate produced. Finally, special commands can be placed into the accompaniment score to force changes in the tempo. These commands allow the system to execute score markings that indicate tempo changes, such as *ritardando* and *A Tempo*. These commands also empty the buffer of position estimates in addition to enacting any specified tempo changes. Thus, the system initially expects the singer to perform the scored tempo changes, but will adjust to the singer's actual performance as new position estimates are made.

The division of computation across two computer platforms requires some form of clock synchronization. Specifically, the observation messages contain time stamps that must be interpreted consistently with the real-time performance of the accompaniment. Work in distributed computing has of course examined clock synchronization issues extensively, and Simons, Welch, and Lynch (1990) provide a good overview of foundational work. However, the case of concern is extremely simple compared to the more complex situations involving networks with many computational nodes and requirements of fault tolerance. Consequently, a very simple solution to the clock synchronization problem could be applied. This solution required a few simple assumptions. Since the only time synchronization requirement relates to the observation messages, the simplest approach is to adjust the time stamps on the messages received by the second computer. Furthermore, it is assumed that there is insignificant latency in delivering a buffer of sampled sound signal to the signal processing code on the first machine. Thus, time stamps generated when a signal buffer is received by this code accurately specify the point in the live performance when the observations were made. Finally, although the clocks on the two computers will drift further and further apart over time, the rate of drift is assumed to be negligible over the period of 5 to 10 seconds. Measurements of this drift for the two computers produced estimates around 0.1 ms/s, scaling to about 1 ms over a period of 10 s. Note that this rather large discrepancy may not be a reflection of differences in clock hardware. The software on one machine receives time from the Windows NT™ operating system while the other uses a simple periodic interrupt.

The time stamp produced by the first machine can be related to time on the second machine as follows:

$$t_2 = t_1 + d_{\text{clock}}$$

$$t_3 = t_1 + d_{\text{clock}} + d_{\text{trans}}$$

where t_1 = Time stamp generated on the first machine.
t_2 = Time on second machine when time stamp was generated on first machine.
t_3 = Time when time stamped message is received by the second machine.
d_{clock} = Time difference between the clocks on the two machines.
d_{trans} = Time difference due to transmission delay.

In order to convert the time stamp in the message to local time, the second machine must obtain an accurate estimate of the time difference between clocks, d_{clock}. A mechanism for providing this estimate was implemented. Specifically, the first machine issues a time stamped message with low transmission delay once every 2 seconds, and the minimum value $t_3 - t_1 = d_{\text{clock}} + d_{\text{trans}}$ over the most recent 3 messages is used as an estimate of d_{clock}. By "low transmission delay" is meant that the time stamp is generated and immediately sent in a distinct message, with no intervening additional computation such as signal processing. Since the message is transmitted via direct MIDI connection between computers and no additional processing occurs on either machine except for sound sample buffering and MIDI playback of the accompaniment, the minimum d_{trans} over three such messages is at most a few milliseconds. Also recall that the drift between the clocks is insignificant over the period of a few seconds, affording use of the minimum difference over three messages spanning several seconds. Reports of observations occur at most once per 100 ms, so the error on the time difference between observations using this synchronization method is one or two orders of magnitude smaller than the time difference itself. The magnitude of the corresponding error introduced to the position estimate is well below the overall accuracy of the tracking system.

The accompaniment control system is based upon a limited model of accompaniment. It reacts to the performer by adjusting both tempo and current position of the accompaniment. It applies a fixed number of alterations based on graded thresholds applied to the difference between the performer's current position and the current accompaniment position. The approach is not terribly detailed, but is not suggested to be a comprehensive model of performances by live accompanists. The model contains two position difference thresholds. Absolute position differences below the first threshold trigger only a tempo change, where the accompaniment tempo is changed to the estimated tempo of the performer. Absolute position differences above the first threshold but below the second result in a compensatory tempo change. If the accompaniment is behind the singer, the accompaniment tempo is increased a fixed percentage above the performer's current tempo. If the accompaniment is ahead of the singer, the accompaniment tempo is reduced a fixed percentage below the performer's current tempo. Thus, the system will either play faster to catch up to the performer or play slower to allow the performer to catch up to the accompaniment. Absolute position differences beyond the second threshold either result in an immediate jump to the position of the performer or cause the accompaniment to stop entirely until the performer catches up.

This approach to controlling the accompaniment performance is based on some obvious behaviors for producing good accompaniment. A study by Mecca (1993) further confirms that human accompanists exhibit certain of these behaviors. First, musical performances inherently contain some amount of motor noise and small timing variation neither controlled nor noticed by the performer. It is important that an accompaniment system not over control the performance or react needlessly to noise. Second, an accompaniment system needs to maintain a smooth and flowing performance, in some instances perhaps more than it needs to achieve precise synchronization with the performer. Sudden jumps, pauses, or hesitations in the melody line of the accompaniment will be noticed and significantly detract from the aesthetics of the performance. Third, drastic control actions such as stopping entirely or jumping ahead several notes should be taken only in the most dire situations. Even in the most critical circumstances, highly competent accompanists may attempt to improvise or modulate between two greatly separated score positions rather than simply jump from one to the other. Typically, reasonable improvisations will be less noticeable and less disturbing than jumps that introduce unmelodic transitions or poor harmonic sequences. Finally, beyond purely aesthetic considerations, applying tempo changes and position difference gradations should help to smooth over errors in score position estimation. In effect, the control system will introduce a hysteresis, gradually resynchronizing based on several successive position estimates rather than instantaneously synchronizing to a single, possibly erroneous estimate.

While the described approach to tempo estimation and accompaniment control is reasonable and has proven viable in real performances, it is possible that accompaniment performance could be enhanced by using more detailed and empirically supported models. Although such investigations are beyond the scope of this work, it would be interesting to develop statistical models for estimating tempo and for deciding accompaniment control actions. Tempo might be estimated by statistical models incorporating the information used to make multiple score position estimates, possibly eliminating intermediate variables for individual position estimates. Accompaniment performance might be based upon more continuous control actions, where tempo changes vary smoothly with score position difference rather than according to coarse gradations. Applying a statistical model to estimate accompaniment control might relax the requirements on tracking accuracy. However, estimating such models probably would require substantial collection and analysis of actual performances, including the accompanists' performances as well as those of the singers. Furthermore, no harm is done by achieving tracking accuracy that is superior to what is ultimately required. The subsequently presented evaluations will remain concerned mainly with the accuracy of score position estimation.

7.3 The Objectives of Evaluation and the Metrics Applied

Evaluation of the score-following model and the accompaniment system assesses both tracking accuracy and synchronization of the performance. However, adequately assessing tracking accuracy is the primary concern. In particular, the evaluation provides summary statistics that assess the expected or average position estimation accuracy of the stochastic score-following model. These statistics will help determine if using a combination of observation types improves tracking accuracy and also the magnitude of any improvement achieved. Also, the evaluation will assess whether actual tracking accuracy is correlated with expected tracking ability based on the expected likelihood of confusing the performer's actual score position with another score position, as defined in the previous chapter.

In addition to average behavior, the evaluation examines extreme outliers. These outliers indicate the most egregious, specific causes of tracking errors. The possible sources of error, as first presented in Chapter 2, are shown in Figure 7-3. Errors due to integral approximation, windowing, and numeric precision were dealt with in Chapters 2 and 3. They will not contribute significantly (if at all) to tracking errors. Sampling along the dimension of score position was also addressed. However, for real-valued observations that are modeled using discrete density functions, sampling along the observation dimension is constrained by the estimation process and the limited available data. Consequently, tracking errors may result from sampling and density estimation, as well as any simplifying assumptions used and any relevant values (variables) that have been omitted. Evaluation of the tracking system should determine at least a possible if not probable cause for extreme outliers, detailing a specific instance of one or more of the identified error categories.

Two metrics are applied to assess the tracking accuracy of the score-following system and the synchronization of the performance. These metrics compare the time stamps on both the estimated score

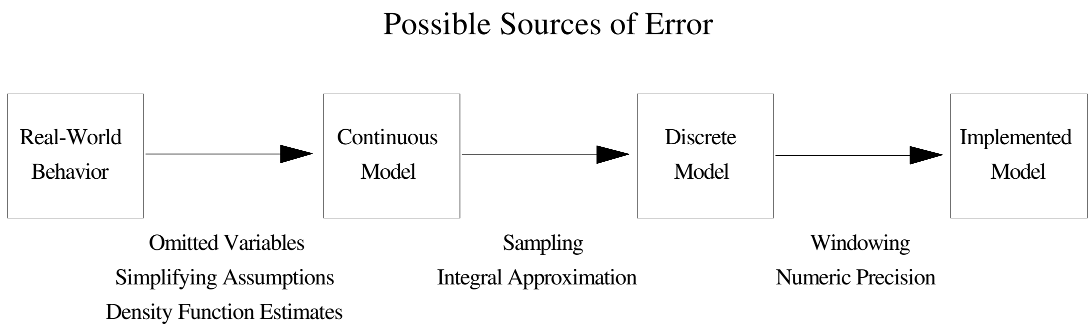


Figure 7-3. Possible sources of error introduced by transitioning between models when implementing the stochastic score-following system.

positions and the notes in the accompaniment performances against human generated segmentations of the soloists' performances. The method of segmenting the performances is identical to that described in Chapter 5. The start time of each note is obtained using a waveform visualization and playback tool. The vowel is taken to indicate the start of a note, except in a melisma where the change in pitch is used. Consistent, steady-state portions of phonemes are easily determined. While the exact point of change from one phoneme to another cannot be identified precisely, a transition region of less than 100 ms can always be identified. In most cases, a clear transition region less than 50 ms is observed. By selecting the center of this region as the start of the later phoneme, the error of the start time estimate is always guaranteed below 50 ms, and is most often guaranteed below 25 ms. In addition, to the extent that one believes it is plausible to attribute the first half of the transition region to the earlier phoneme and the second half to the later phoneme, the actual error in all cases is below 25 ms. Changes in pitch within a melisma can be identified with similar accuracy.

During accompaniment of each performance, all position estimates made by the tracking system are time stamped and recorded. The recorded times are the times contained in the observation messages. Time differences are calculated for the first position estimate within each note in a given score:

$$\text{TrackingTimeDifference} = \text{Time System Estimated Position} - \text{Time Singer Performed Position}$$

Time stamps also are recorded for notes performed during accompaniment playback:

$$\text{SynchronizationTimeDifference} = \text{Time System Performed Position} - \text{Time Singer Performed Position}$$

When applying this metric, time differences are calculated only for each note in the accompanist's part that is scored for simultaneous performance with a note in the soloist's part. In the case of chords and intervals, only one time difference is calculated based on the earliest time stamp of any note in the chord or interval. Time stamped sequences are aligned with hand segmentations by identifying a reliable synchronization point—generally a note that follows a rest and begins with significant amplitude. The calculated time differences are used to generate summary statistics for individual performances and across sets of performances. For different performances of the same score, or different trials of the same recording, the time differences can also be compared on a per note basis.

Finally, it is worthwhile to consider the time difference statistics compared against the available knowledge on human perception of music and human musical performance ability. However, as of the time of this study, limited published assessments of human performance on comparable tasks are available. Mecca (1993) examined deviations in synchronization for a human accompanist playing against a computer generated metronomic pulse. The standard deviations of time differences produced by individual accompanists ranged from 5 to 48 ms. Several studies have examined timing deviations of individual notes in performed melodies and rhythmic patterns. Perception studies have shown that human listeners can sometimes discern deviations in the range of 20 to 50 ms, and are likely to discern deviations above 50 ms (Clarke 1989; Schulze 1989). In contrast, studies of human performances commonly report that timing deviations due to motor noise (unintentional deviations) range from 10 to 100 ms (Desain and Honing 1992), and timing deviations in rhythmic sequences reproduced by tapping show skilled subjects produce standard deviations around 100 to 140 ms (Povel and Essens 1985). Skilled musicians who were asked to listen to a rhythmic sequence and then to reproduce a specified fraction of a beat by tapping exhibit timing deviations with average standard deviations between 25 and 60 ms (Sternberg, Knoll, and Zukofsky 1982).

While these numbers are informative, comparing such numbers against tracking system accuracy must be done cautiously. The metrics used to assess musical performance and perception are overly general at best. Timing deviations in performances and the perception of those differences depend upon compositional style, performance style, and the musical context surrounding individual notes. A timing difference of 100 ms during a rapid melisma in a Handel aria may not be perceived in the same way as a timing difference of 100 ms during a slow rubato passage in a German lied. Also, assessing the tracking system based on unrealistic, conservative timing requirements may lead to overly pessimistic conclusions, essentially demanding superhuman listening ability from the accompaniment system. The difficulty and complexity of defining appropriate metrics for assessing music systems is a recognized challenge in computer music (Hirata 1997). In addition, the applied approach to controlling accompaniment performance is not claimed to be comprehensive and detailed, and undoubtedly a more thorough exploration of human accompaniment could lead to enhanced computer performance as well as improved metrics. Finally, nearly all performance timing experiments have dealt with musical instruments or computer generated sounds. Perception of singing, performance timing in singing, and synchronization between singers and accompanists may exhibit different characteristics. Since humans perceive the lyrics as well as the melody in singing, and pitch in singing is much more variable, larger differences in expressive timing and synchronization may be tolerated or more difficult to recognize.

The interaction between perception and performance in musical accompaniment is certainly very complex. Differences between what is easy for a computer and what is easy for a human further complicate comparisons. For instance, although humans can perceive timing differences below 50 ms, reaction time to an auditory stimulus is between 100 to 150 ms. These numbers imply that humans must plan for synchronized performances, relying on an internal sense of time and anticipating note production in order to bring synchronization errors below 100 ms. The developed accompaniment system, however, estimates position with only a 50 ms latency and can perform notes in just a few milliseconds after completing position estimation. Using contemporary processors with higher clock speeds, this total turnaround time can be reduced even further. Thus, highly accurate tracking may be able to compensate for some lack of performance planning. On the other hand, improved performance planning might compensate for inaccurate tracking. In some instances, this planning may still be necessary in order to guarantee acceptable synchronization, either because the total performance latency is still too great or because the sound signal does not provide sufficient information to discern actual position. To the extent that humans rely on performance planning and expectations, the information provided by the sound signal may be less relevant. While not directly addressed in this work, consideration of these tradeoffs and the optimization of reliance on tracking accuracy versus performance planning is worthy of subsequent investigation.

7.4 Evaluation Using Recorded Performances

To evaluate the accompaniment system, two sets of performance trials were analyzed. The first set of trials used recorded performances while the second set used live performances. Performing with recordings can be viewed as a "stress test" for both tracking and accompaniment, since the recording will never adjust to compensate for errors in the computer performance. A set of eight recordings were used in the performance trials. These recordings are of trained vocalists performing Western classical music, all vocalists having at least one year of university training. The eight recordings were not used to estimate the density functions in the stochastic score-following model. The recordings contained four performances by female singers and four by male singers, with representation from all primary voice parts. Two of the performances were given by singers who were not represented in the performances analyzed for density function estimation. If a singer was represented in both sets of performances, the two performances were of different pieces. The pieces performed spanned a variety of styles and genres, including examples of operatic arias, lieder, and arrangements of folk songs. The pieces contained lyrics in either Italian, German, or English.

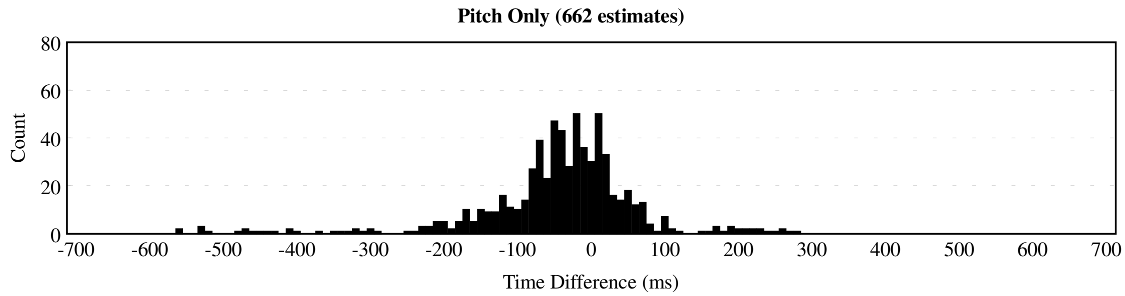
Two forms of scores were prepared for each piece in the recording. The first was a performance score that contained both the soloist and accompaniment parts. This score specified pitch and rhythm (relative duration) for each note, dynamic changes, and nominal tempi for each section of the piece. In addition, special tempo or duration information was provided for five special case scenarios. First, commands are placed at the beginning of extended rests to indicate that the accompaniment should proceed at the current estimated tempo. This prevents the system from pausing or slowing if the singer extends the final note of the phrase into the rest. A complementary command placed before the end of each extended rest indicates that the accompaniment should resume following the singer at the next entrance of the vocal part. Second, commands are placed near the end of each extended rest to force a reset of the buffer containing time stamped position estimates. These commands preclude tempo estimates based on position estimates that span the rest. In the case of recordings, the human accompanist controls the tempo during these rests, and the tempo may change when the singer enters. Averaging across a large rest is not likely to produce a good estimate of expected tempo. Third, when an *A Tempo* (translates "return to tempo") marking appears, a pair of commands is placed in the score to record tempo at a point preceding the tempo change (at least 1 full measure before the marking) and reset tempo to the recorded tempo at the point where the *A Tempo* marking occurs. These commands automatically reset the buffer of time stamped position estimates. Fourth, a similar mechanism is applied whenever a cadenza or *espressivo* (translates "expressively") section appears. In addition to recording and restoring the tempo, however, the expected tempo is reduced by twenty percent at the beginning of these sections. This adjustment prevents the convolution process from overwhelming the observation distributions and producing erroneous position estimates based on an estimated tempo that is much too fast. Finally, a fermata placed over a note or rest causes the notated relative duration to be extended by either a factor of two or one quarter note, whichever yields the longer duration. In the case of a fermata placed over a note, the total extended duration of the note is then reduced in order to insert an eighth-rest after the note.

While some of the explicit tempo and duration alterations affect only accompaniment control, such as the "ignore singer" markings, others affect position and tempo estimation. Obviously these explicit alterations are heuristic. A more comprehensive and thorough approach would be to include the corresponding score markings as explicit conditioning variables in the distance density function and estimate an alternative model. For instance, a variable could be included to indicate presence of or proximity to an *espressivo* marking, and estimates could be generated for the appropriate distance density under all circumstances. While the resulting distance density function would be more comprehensive, it requires many examples of *espressivo* sections. Since only a limited number of performances could be analyzed for this study, such an approach was not possible. Consequently, the described heuristic approach was taken. However, the specifics of the tempo and duration adjustments were based upon the few examples available in the examined performances, in addition to the experiences of the author.

The second score prepared for each performance contained all the information for specifying the events in the stochastic score-following model. This information included all pitches, durations, and tempi that appeared in the vocalist's part in the first score. In addition, the second score contained a phonetic transcription of the lyrics in the piece, along with start times that aligned the vowel in each syllable of the lyrics with a pitch in the first score. Transcriptions were generated directly from the score and not from examination of the recordings. Events based on fundamental pitch, spectral envelope, and note onsets were generated from this information. Details of this process were previously described for each respective observation type.

Given the two scores, the accompaniment system could be used to track each recording and provide real-time accompaniment. Prior to executing the trials for each recording, sound levels were adjusted on all equipment and an amplitude threshold determined. Sound levels were set to maximize the input level but avoid clipping of the signal. Threshold settings were determined as previously described. Each recording was played several times and a threshold selected so that the softest note in the recording still triggered pitch detection. The minimum time span for estimating tempo was set at 1 second and the preferred time span at 3 seconds. These settings enabled the system to use a more stable estimate when available but still use a recent estimate after the position buffer was emptied. The two absolute position difference thresholds for determining corrective performance actions were set at 50 ms and 750 ms. Corrective tempo adjustments increased or decreased the accompaniment tempo by 10% relative to the estimated tempo of the singer. Several trials were run for each recording, and time stamped position estimates and performance traces were generated. The experiment included three trials per recording for each of five different sets of observation types—pitch alone, spectral envelope alone, pitch and spectral envelope, pitch and note onsets, and all three observation types. Handmade segmentations of the recorded performances were prepared, and time differences between the recorded data and the handmade segmentations were generated for each trial. Graphs and summary statistics over the time difference data were subsequently assessed and compared.

First considered were the differences between the time a score position was estimated and the time the singer was actually at that position. Several general properties of the distribution of these time differences are observed in graphs of the differences. Figure 7-4 shows two histograms of the time differences generated for a single recording. The upper histogram includes data from the three trials using only pitch observations and the lower graph displays data from the three trials using all observation types.



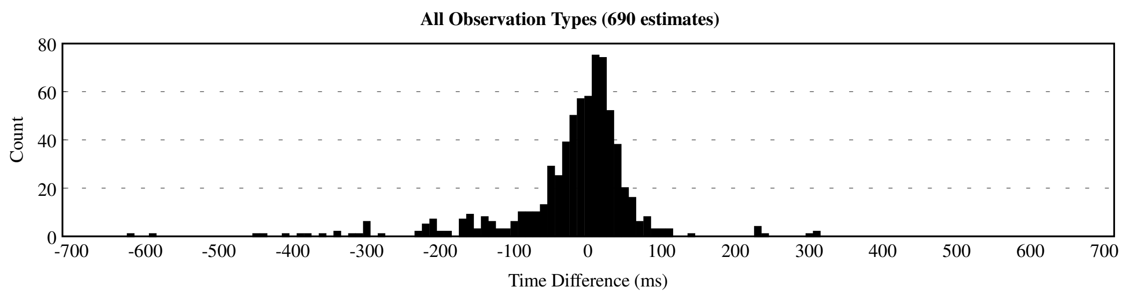


Figure 7-4. Histograms of the differences between the time a position was estimated and the time the performer was actually at that position. Positive values indicate the estimate was late, negative values indicate the estimate was early. Each graph includes time differences from all three trials of one recording when using the specified observation types.

Positive values indicate the tracking system estimated the position too late, and negative values indicate the system estimated the position too early. Note that in both cases, most estimated positions deviate from the actual positions by less than 100 ms. Also, the distribution sharpens when all observation types are used to track the performer. This property is typical of the graphs for all recordings. Although the tails in the graph narrow when all observation types are used, there remain some extreme outliers beyond the outliers appearing when only fundamental pitch is observed. Unfortunately, this situation is also commonly seen in the graphs.

Summary statistics for each performance are presented in Table 7-1. Standard deviations and means were calculated for the time differences in each individual trial and averaged to produce the given results. Most of the means are negative, possibly indicating a tendency for the system to estimate a position slightly early rather than slightly late. Since the lognormal distribution is positive valued, the tracking system may be biased toward confusing the actual position with a succeeding position rather than with a preceding position. Also, it is arguable that most significant, unwritten tempo alterations involve an extended, significant slowing followed by an immediate return to the original tempo. For instance, this situation occurs when a *ritardando* is applied at the end of a phrase. Thus, music performances may provide greater opportunity to overestimate the tempo, resulting in more negative outliers than positive

Table 7-1. Mean and standard deviation of time differences for recorded performances. The values for each performance were averaged over three trials for each set of observation types.

	Pitch	Spectral Envelope	Pitch & Spectral Envelope	Pitch & Onsets	All Observations
Performance 1					
Mean	-58 ms	-135 ms	-70 ms	-96 ms	-23 ms
SD	208 ms	374 ms	203 ms	288 ms	175 ms
Performance 2					
Mean	-23 ms	68 ms	-20 ms	14 ms	-7 ms
SD	94 ms	431 ms	99 ms	87 ms	60 ms
Performance 3					
Mean	-13 ms	793 ms	-24 ms	-24 ms	-18 ms
SD	132 ms	1570 ms	176 ms	131 ms	154 ms
Performance 4					
Mean	-144 ms	———	-125 ms	-166 ms	-147 ms
SD	278 ms	———	323 ms	299 ms	355 ms
Performance 5					
Mean	-97 ms	-80 ms	-90 ms	-72 ms	-69 ms
SD	292 ms	298 ms	186 ms	243 ms	200 ms
Performance 6					
Mean	-49 ms	-41 ms	-33 ms	-34 ms	-25 ms
SD	111 ms	196 ms	117 ms	113 ms	90 ms
Performance 7					
Mean	-13 ms	-356 ms	-14 ms	-28 ms	-21 ms
SD	138 ms	2091 ms	114 ms	131 ms	94 ms
Performance 8					
Mean	-30 ms	1307 ms	-35 ms	-58 ms	-57 ms
SD	151 ms	2222 ms	179 ms	138 ms	147 ms
Average					
Mean	-53 ms	222 ms	-51 ms	-58 ms	-46 ms
SD	175 ms	1026 ms	175 ms	179 ms	159 ms

outliers. The histograms of time differences for certain performances confirm this situation. Finally, errors in time aligning the estimated positions with the hand segmentations could produce negative means. Adding an initial short duration click on the recordings could alleviate this concern if the signal processing system were designed to observe the click and record a time stamp for it.

Standard deviations for the individual performances appear large. In particular, observing only spectral envelope is extremely ineffective. Trials using this version of the tracking system were very poor. In the case of performance four, the system could not follow the singer through the piece, becoming hopelessly lost before completion of the recording. Given the previously expressed concerns over the estimated distributions for spectral envelope and the omitted conditioning variables, it is not surprising that this version of the system performs poorly. More surprising is the lack of significant improvements when observing multiple observations relative to observing only fundamental pitch. Trials using the pitch only version of the tracking system were good overall. In all trials, the system was able to follow the singer through the entire piece. Audible synchronization problems occurred mainly in instances when the tempo was altered suddenly and significantly. Using all observations, the standard deviation was reduced for six of the performances. A paired comparisons sign test (a test considering how frequently 6 out of 8 trials would show improvement if the two systems were identical) gives a P-value just under 0.15 for this situation. The average standard deviation drops from 175 to 159. However, in the case where all observation types were used, the average is strongly affected by the large increase in the standard deviation for performance four. This increase is due mainly to a few extreme outliers that occurred within a cadenza section in the performance—a section where the accompanist does not play. Standard deviations for the case where pitch and spectral envelope are used and the case where pitch and note onsets are used show little improvement over the case where only pitch is used.

Since the mean and especially the standard deviation are affected by significant outliers, these statistics were calculated over the time differences when the most significant 5% outliers by absolute time difference were discarded. Table 7-2 provides the summary statistics for the adjusted data sets. Naturally the means and standard deviations are smaller. However, note the improvement for relative comparisons of standard deviations between the various versions of the tracking system. Improved tracking resulted for all eight performances when using all observation types compared to using only pitch, and for seven performances when using pitch and onsets compared to using only pitch. Paired comparisons tests for these cases yield P-values of 0.004 and 0.035 respectively. Tracking improved for five performances when observing both pitch and spectral envelope compared to observing only pitch, and the remaining three performances have standard deviations that are within 11 ms of one another. Overall, the average standard deviation decreases as the number of observation types increases. The average standard deviation when using all observation types is 30% less than the average standard deviation when observing only fundamental pitch. Note that the standard deviations in the two tables support the original statement made about the histograms of time differences, as in Figure 7-4. Namely, the center regions of the histograms narrow as the number of observation types increases, but some large outliers remain.

The standard deviations calculated per performance are influenced by significant outliers and are averaged over a limited number of performances. As another approach to comparing tracking accuracy between versions of the score-following system, paired comparisons tests were run based on pairing position estimates for the same notes in the score. Specifically, for each performance, each trial using only pitch was paired with a trial using all observation types. For each note in the score, the first position

Table 7-2. Mean and standard deviation of time differences for recorded performances with 5% outliers discarded. The values for each performance were averaged over three trials for each set of observation types. Outliers were removed per trial, based on absolute time differences.

	Pitch	Spectral Envelope	Pitch & Spectral Envelope	Pitch & Onsets	All Observations
Performance 1					
Mean	-38 ms	-73 ms	-31 ms	-72 ms	-10 ms
SD	155 ms	230 ms	121 ms	194 ms	96 ms
Performance 2					
Mean	-8 ms	79 ms	-4 ms	22 ms	0 ms
SD	52 ms	335 ms	59 ms	36 ms	34 ms
Performance 3					
Mean	5 ms	520 ms	6 ms	-4 ms	11 ms
SD	97 ms	1069 ms	108 ms	68 ms	76 ms
Performance 4					
Mean	-122 ms	———	-119 ms	-134 ms	-130 ms
SD	209 ms	———	195 ms	194 ms	170 ms
Performance 5					
Mean	-47 ms	-26 ms	-64 ms	-27 ms	-28 ms
SD	159 ms	146 ms	89 ms	128 ms	72 ms
Performance 6					
Mean	-35 ms	-7 ms	-15 ms	-23 ms	-16 ms
SD	75 ms	88 ms	67 ms	63 ms	56 ms
Performance 7					
Mean	-13 ms	-126 ms	-8 ms	-16 ms	-9 ms
SD	85 ms	1814 ms	65 ms	66 ms	60 ms
Performance 8					
Mean	-12 ms	927 ms	-18 ms	-33 ms	-31 ms
SD	108 ms	1688 ms	117 ms	71 ms	78 ms
Average					
Mean	-34 ms	185 ms	-32 ms	-36 ms	-27 ms
SD	117 ms	767 ms	103 ms	103 ms	80 ms

estimates within the note were paired across the two trials, and the differences of the differences calculated. T-tests were run for each set of differences of differences based on a null hypothesis that the mean is zero. Results of these tests are presented in Table 7-3. Seventeen of twenty-four tests showed improved tracking using all observations with a P-value less than 0.15. Only one performance provided no paired trials that showed improvement at this level—performance four. However, no tests showed degraded performance at this level when using all observations. The average of the mean differences over all pairs of trials was -22 ms. Excluding performance four, this average was -25 ms.

Table 7-3. Results of paired comparisons tests when pairing estimated positions for the same note across two trials. Each test considered two trials of the same performance, one trial using only pitch and one using all observation types. Three tests were run for each of eight performances, for a total of 24 tests.

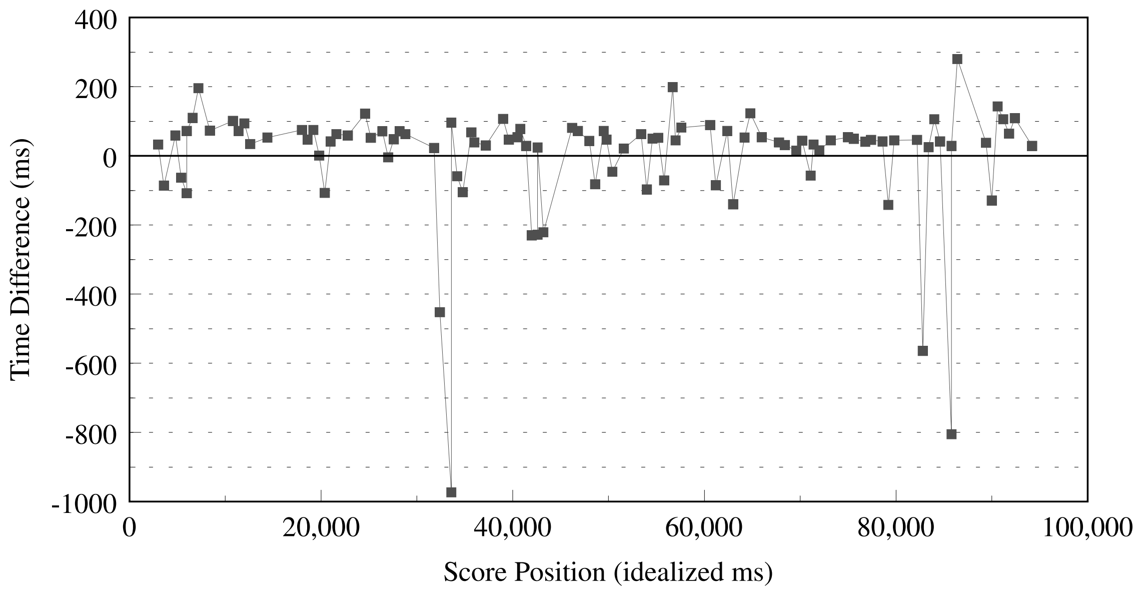
P-Value per Paired Comparisons Test	Tests Showing Improvement Using All Observations	Tests Showing Degradation Using All Observations
<0.01	9	0
<0.05	14	0
<0.10	16	0
<0.15	17	0

Although large time differences sometimes appear as outliers, these position estimation errors may not be serious for accompaniment if they are not correlated or they occur during unaccompanied portions of the vocal performance (such as cadenzas). Graphs of the time differences over the duration of individual pieces were examined. These graphs commonly revealed little proximity between large time differences, as shown by the upper graph in Figure 7-5. Sometimes a few of the outliers were clustered, appearing during cadenzas or unaccompanied sections in a recitativo. The second graph in Figure 7-5 shows time differences for a trial of performance one. Note that the large outliers appear near the beginning and end of the performance. These sections of the piece were performed more freely and contained minimal accompaniment.

In addition to accuracy of position estimation, synchronization of the accompaniment was also considered. Specifically, for every note in the accompaniment that was scored to coincide with a note in the vocalist's part, the time difference between the singer's performance and the computer performance was examined. Figure 7-6 shows two histograms of the time differences generated for a single recording. The upper histogram shows time differences in synchronization and the lower graph displays time differences in tracking. Positive values in the second graph indicate the tracking system estimated the position too late, and negative values indicate the system estimated the position too early. Similarly, positive values in the first graph indicate the system performed notes at the given score position after the singer performed that position, and negative values indicate the system performance was early.

Two important properties of the synchronization time differences are observed in the graphs. First, some of the synchronization graphs contain fewer or less extreme outliers than the tracking graphs, while some exhibit more outliers. This situation occurs because of the hysteresis introduced by the accompaniment control system relative to the position estimates. The muted reaction to position estimates sometimes helps performance by preventing the system from immediately synchronizing with inaccurate

a) Trial of performance three using all observation types:



b) Trial of performance one using all observation types:

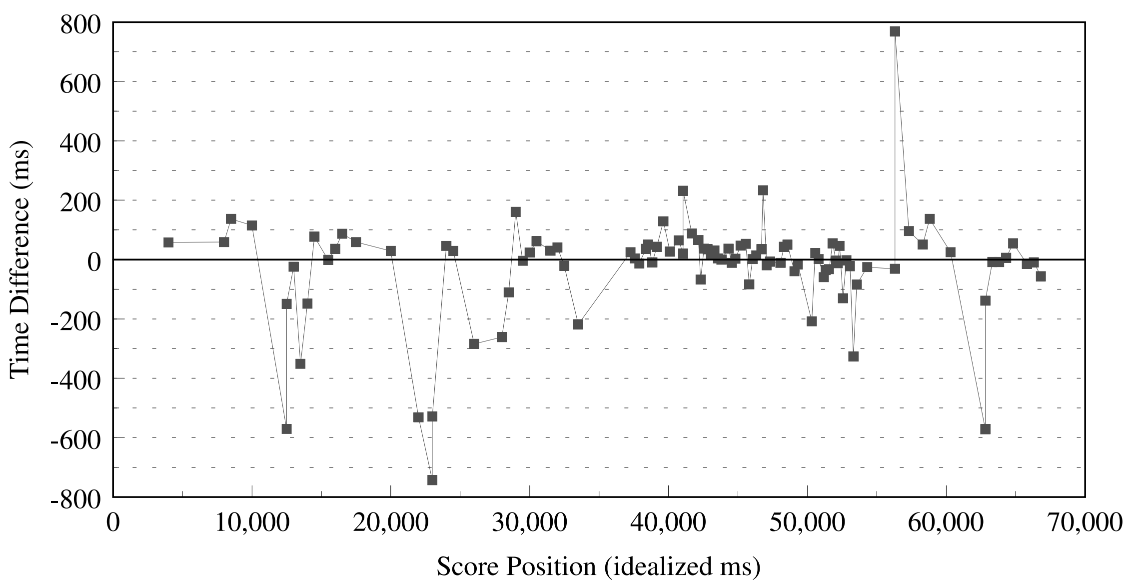


Figure 7-5. Time differences for estimated score positions in two trials with recorded performances. Graph a) shows time differences for a trial of performance three, showing little correlation between point in the performance (score position) and extreme time differences. Graph b) shows time differences for a trial of performance one, showing extreme time differences near the beginning and end of the piece. The first and last sections in the song are performed more freely and contain little accompaniment.

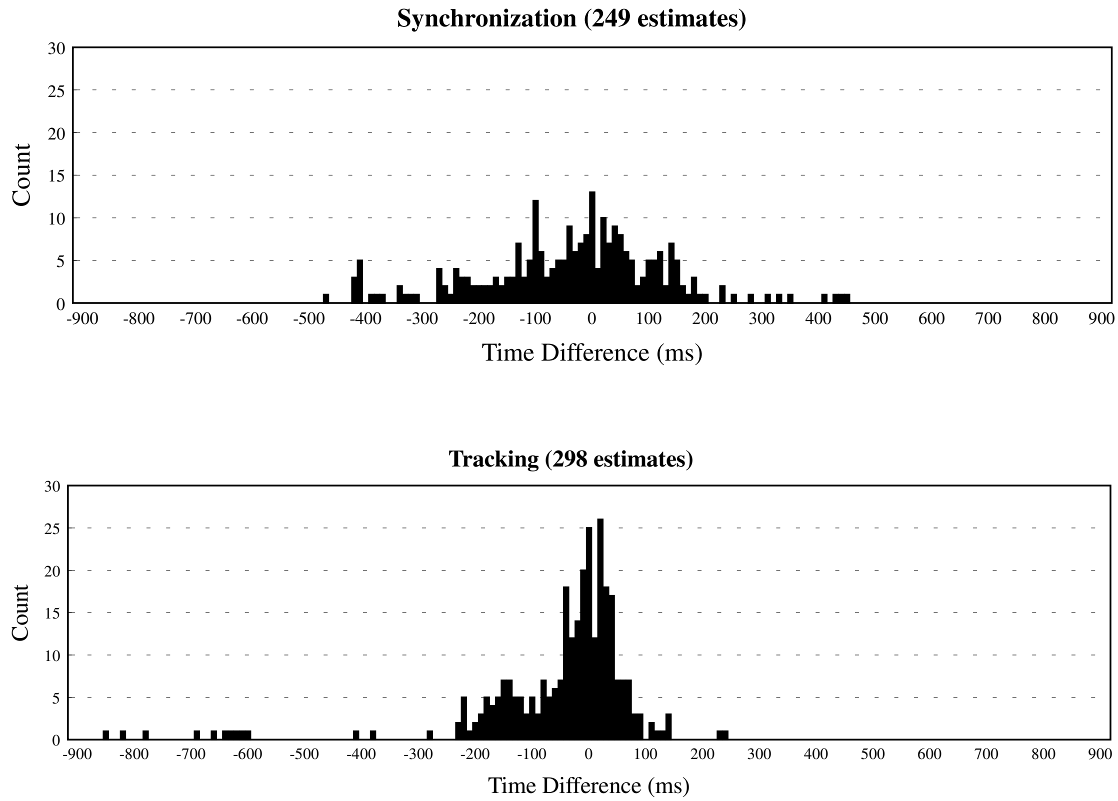


Figure 7-6. Histograms of synchronization and tracking time differences for three trials of the same performance, using all observation types. Time differences for synchronization indicate the time the accompaniment system performed notes at a given score position minus the time the singer performed that position. Time differences for tracking indicate the time the system estimated a score position minus the time the singer performed that position.

estimates, and sometimes hinders performance by preventing the system from immediately responding to significant tempo changes. Second, the synchronization graphs are generally less sharp than the tracking graphs. Several possible factors might produce this characteristic. The hysteresis introduced by the accompaniment control system could increase individual time differences while maintaining an average close to zero. Also, the system latency might increase the error in the synchronization actions relative to the error of the position estimates. However, this latency is small, so it is less likely to account for the total observed increase in the spread than the hysteresis resulting from accompaniment control. Finally, all time differences have been calculated assuming that the vowel in each syllable represents the start of the note. If sometimes the singer actually intends any preceding consonants to coincide with the notes in the accompaniment, the accompaniment system may perform notes correctly but will be penalized by the metric. The loose coupling between position estimation and accompaniment performance can allow the system to perform notes slightly early relative to the estimate of the singer's position, permitting the tracking system to estimate the start of the vowel with higher accuracy.

Summary statistics for each performance are presented in Table 7-4. Standard deviations and means were calculated for the time differences in each individual trial and averaged to produce the given results. Overall, there are not as many negative means for these averages as for the averages over tracking time differences. The accompaniment control system may limit effects of any bias toward premature estimation of positions exhibited by the tracking system. The standard deviations are often larger for the synchronization time differences than for the tracking time differences. This effect is expected given the previous discussion of the graphs. In most of the cases where standard deviation is smaller for the synchronization time differences, the standard deviation for the tracking time differences is large relative to the other performances (normally near or above 200 ms). These numbers probably indicate that the

Table 7-4. Mean and standard deviation of synchronization time differences for recorded performances. The values for each performance were averaged over three trials for each set of observation types.

	Pitch	Pitch & Spectral Envelope	Pitch & Onsets	All Observations
Performance 1				
Mean	137 ms	84 ms	29 ms	89 ms
SD	160 ms	184 ms	229 ms	186 ms
Performance 2				
Mean	2 ms	5 ms	54 ms	33 ms
SD	129 ms	136 ms	146 ms	130 ms
Performance 3				
Mean	15 ms	-24 ms	-25 ms	1 ms
SD	157 ms	207 ms	142 ms	153 ms
Performance 4				
Mean	-131 ms	-156 ms	-125 ms	-177 ms
SD	302 ms	324 ms	267 ms	405 ms
Performance 5				
Mean	-5 ms	17 ms	1 ms	14 ms
SD	256 ms	224 ms	247 ms	188 ms
Performance 6				
Mean	-28 ms	4 ms	23 ms	14 ms
SD	137 ms	171 ms	188 ms	161 ms
Performance 7				
Mean	-71 ms	-62 ms	-79 ms	-65 ms
SD	179 ms	195 ms	168 ms	173 ms
Performance 8				
Mean	-22 ms	-33 ms	-38 ms	-38 ms
SD	162 ms	219 ms	139 ms	161 ms
Average				
Mean	-13 ms	-21 ms	-20 ms	-16 ms
SD	185 ms	207 ms	191 ms	195 ms

accompaniment control system can improve synchronization when large position estimation errors occur, but that the resulting hysteresis hinders synchronization when the estimates are accurate. Based on the presented standard deviations, using multiple observations for performance tracking does not appear to improve synchronization on average. Removal of 5% outliers by absolute value does not support a different conclusion. The given accompaniment control system does not leverage any benefits from the improved position estimation.

Finally, noticeable performance errors are sometimes (but not always) correlated with extreme errors in position estimation. It is important to identify the primary causes of extreme estimation errors. Subsequent work on performance tracking could focus on eliminating these errors, thus improving tracking accuracy and reducing the need for accompaniment control that introduces hysteresis. To provide some insight into the most significant errors, time differences in the trials using all observations were examined. All tracking time differences with an absolute value beyond 300 ms were attributed to one of four categories. With the exception of performance four, values this extreme constituted fewer than 8% of all time differences calculated for each performance. A total of 271 extreme time differences were examined. The categories and the distribution of the extreme time differences appears in Table 7-5. An attempt was made to attribute each difference to a single category. However, in cases where large time differences were caused by multiple effects spanning two or more categories, a fraction was added to the counts for each category.

The categories represent a general partitioning of the position estimation "errors". "Threshold/Detector Error" indicates that an inappropriate portion of sound signal was processed, resulting in invalid pitch and spectral envelope observations or a spurious onset. For instance, one consonant in a cluster might have been louder than the others, and its amplitude exceeded the threshold. "Added Rest or Breath" indicates errors that resulted because the singer inserted a rest or breath where

Table 7-5. Causes of extreme time differences and the percent of the extreme differences in the trials using all observations that are attributed to each cause. A total of 271 time differences were considered.

Causes of Extreme Time Differences	Percent of the Extreme Time Differences (absolute value > 300 ms)
Threshold/Detector Error	22%
Added Rest or Breath	13%
Altered Duration or Tempo	32%
Low Probability for Observations	35%

none was marked in the score. This action often triggers an onset where none is expected and sometimes delays the subsequent note, effectively invalidating the applied model of a singer's motion. "Altered Duration or Tempo" indicates errors that result from unexpected tempo changes or lengthening of notes and rests. Finally, "Low Probability for Observations" indicates errors due to attributing more likelihood to an incorrect score position based on the observations. These errors cannot be explained other than by observation density functions that do not contain all relevant conditioning variables or that have been estimated inaccurately.

In terms of the possible sources of error described in Section 7.3, the large time differences are caused by a combination of omitted variables, simplifying assumptions, and inaccurate estimation of certain density functions. For instance, problems with the amplitude threshold result from omitting relevant information determining amplitude, as described in Chapter 6. However, a model including all this information would be challenging to estimate, even if such a comprehensive model could be defined. A better solution might be to use an alternative method of distinguishing pitched from unpitched signal, possibly redefining the onset events. Another prominent source of errors, the possibility that the singer will insert a rest or breath, is not considered in the model. To deal with this problem, the modeling of events or the score must be extended. Such modifications are likely to influence both the definition and estimation of the distance density and the observation densities. Altered durations and tempi constitute a previously anticipated problem. As already recognized, the model of motion presented in Chapter 3 only modeled average behavior, and the outliers did appear in the fitted convolution model. Extending the model of motion to account for large tempo changes could help, though producing such an enhanced model would require substantial work and tedious data collection. Also, improving the observation densities or adding observation types may help to counteract errors introduced by the model of motion. Finally, low probability observations occur most often as unanticipated spectral envelope values, and this situation was expected. Cases of unlikely observed fundamental pitch are less common but do occur, particularly when the singer is altering the pitch for expressive purposes. Examples include applying portamento and ornamenting the performance. Enhanced observation distributions, particularly for spectral envelope, would improve tracking accuracy.

The accompaniment control system uses position estimates that correspond to the center of the 100 ms region of score most likely to encompass the singer's current position, according to the score position density function. During trials of each performance, two other position estimates also were calculated and recorded. These estimates included the mean of the score position density (the estimate that minimizes the error variance) and the mode of the score position density (the maximum of the sampled points). Table 7-6 contains standard deviations of the tracking time differences when using each point estimate of the singer's score position. These statistics were calculated over the trials using all observation types. Note that overall the three position estimates yield comparable results with respect to the time difference metric. These results indicate that the score position density generally contains a significant amount of area in a single, relatively small region of the score, this region containing the mode of the function. It is less common that the score position density contains multiple, distinctly separated "bumps" each containing significant area—a situation that would cause the mean estimate to deviate from the mode and maximum area estimates.

When all data points are considered, the mean outperforms the other two estimates in 6 out of 8 performances. When the 5% outliers are discarded for each estimate, the maximum area estimate outperforms the mean estimate in 6 out of 8 performances and the mode estimate in 7 out of 8 performances. This behavior might be explained as follows. In situations where distinct regions of the score each contain significant area (*i.e.*, the stochastic model does not clearly discriminate a single note), the mean estimate will produce a score position somewhere between the two regions while the maximum area estimate will produce a position within the region containing the most area. Sometimes the maximum area estimate will select the region farthest from the performer's actual position and consequently will generate extreme outliers that are larger than the outliers produced by the mean estimate. Thus the mean estimate will produce a time difference distribution with smaller variance, even though synchronizing an accompaniment to a score position between the two most likely regions may

Table 7-6. Standard deviation of tracking time differences for recorded performances when using different point estimates of score position. The values for each performance were averaged over the three trials when using all observation types. Outliers are discarded based on absolute time difference.

	All Data			5% Outliers Removed		
	Max. Area	Mean	Mode	Max. Area	Mean	Mode
Performance 1	175 ms	198 ms	164 ms	96 ms	109 ms	100 ms
Performance 2	60 ms	53 ms	68 ms	34 ms	36 ms	43 ms
Performance 3	154 ms	126 ms	173 ms	76 ms	79 ms	94 ms
Performance 4	355 ms	337 ms	367 ms	170 ms	163 ms	179 ms
Performance 5	200 ms	194 ms	200 ms	72 ms	75 ms	86 ms
Performance 6	90 ms	90 ms	90 ms	56 ms	58 ms	60 ms
Performance 7	94 ms	84 ms	94 ms	60 ms	61 ms	61 ms
Performance 8	147 ms	141 ms	153 ms	78 ms	76 ms	75 ms
Average	159 ms	153 ms	164 ms	80 ms	82 ms	87 ms

make no sense musically. However, when the density function contains only a single region of prominent area encompassing the actual position of the singer, the mean estimate may be pulled off center by other less prominent but nontrivial "bumps". Thus when the extreme outliers are discarded, the maximum area estimate may have an advantage since it essentially ignores area outside of a small region local to the most prominent bump. If the stochastic model is improved and the number of significant time difference outliers reduced, essentially producing sharper score position densities on average, then the maximum area estimate may be preferred over the mean estimate. However, probably both are acceptable for accompaniment control. The mode estimate is probably less effective due to the skew of the score position density, resulting from convolution by a lognormal density function.

In summary, accompaniment trials involving recorded performances have provided useful information about the accuracy of the tracking system. First, fundamental pitch is the best choice for a single observation type because it is informative, relatively reliable, and easy to model accurately. Second, combining observation types does produce improved tracking, both on a per performance basis and on average. However, significant outliers can result from unexpected, significant tempo changes and poor estimation of the observation densities. The outliers often can be addressed by modifying the scores, if not too many occur. Furthermore, some tracking outliers are unimportant for accompaniment if they occur in sections of the piece that are unaccompanied or sparsely accompanied, including cadenzas and recitativos. The accompaniment control system also introduces a hysteresis, so if the outliers do not occur in clusters they may not be noticeable in a performance. The degree of hysteresis introduced by a simple approach to accompaniment control prevents the accompaniment system from taking advantage of the improved position estimation obtained by using multiple observation types. It would be useful to devise a means for optimizing the accompaniment control system based on statistics for a given tracking system's accuracy. Overall, the performance tracking system does not rival human perception, but with low latency and effective accompaniment control, the combined system can produce reasonable accompaniment in many instances.

7.5 Evaluation Using Live Performances

The second set of analyzed performance trials was comprised of live performances. A set of six performances was examined. These performances were given by trained vocalists (all of them students majoring in vocal performance) singing Western classical music. These vocalists had at least one year of university training. Two of the performances were by male singers and four by female singers. Represented voice parts included soprano, mezzo, tenor, and baritone. Only one of the singers had provided recordings that were used for density function estimation. The performed pieces spanned a variety of styles and genres, including examples of operatic arias, lieder, and contemporary art song. The pieces contained lyrics in either Italian, German, or English. The performers were asked to select a song of their choice from their repertoire in one of these languages.

As with the trials using recordings, two scores were prepared for each performance. The scores were prepared as previously described with the exception that the singers were asked to provide their own phonetic transcriptions of the lyrics. Before performing each trial, the process of the experiment was explained to the singer. The singers were told that the automated accompaniment system had the score and would listen to them through a microphone on a stand. The system was designed to follow them and adjust tempo as they did. The purpose of the experiment was to assess how accurately the system tracked their position in the score. Their performances would be recorded. Prior to recording trials for analysis purposes, some sound levels would be adjusted and a few trial performances attempted to insure that the system was properly configured and functioning. The singers were not told in advance that the recorded trials would involve different configurations of the score-following system, observing different aspects of the performance.

The accompaniment part was sounded through a set of four speakers positioned in the corners of the room. The singers were positioned equidistant from all four speakers with the microphone slightly in front of them. This positioning helped to equalize the sound in both of the singer's ears. Since the floor was carpeted and the walls tiled with soft panels, the room contributed little reverberation. Prior to any performances, sound levels were adjusted on all equipment and an amplitude threshold determined. To set levels on the mixer and sound card, the singers were asked to perform excerpts from the loudest and softest portions of the piece. Sound levels were set to maximize the dynamic range but avoid clipping of the signal. Threshold settings were determined by setting the threshold as high as possible while still triggering pitch detection during the softest excerpt.

Once level settings were configured, the accompaniment was played through the speakers. The signal processing system was operated during this playback, in order to verify that the accompaniment would not trigger observations while the singer was silent. In addition, the singer was asked to listen to the accompaniment, both to become familiar with the sound and to comment on the initial tempo. Since most vocal pieces begin with an accompaniment introduction, the accompanist establishes the initial tempo. Nominal tempi in the scores were invariably too slow. This result was expected since slow tempi were selected deliberately to insure accuracy of the numerical convolution. Also, negotiation of initial tempo between singers and human accompanists is not uncommon preceding their first joint performance of a piece. The initial tempo was adjusted upwards until the singer expressed satisfaction with the playback of the accompaniment introduction.

After setting sound levels and initial tempo, a performance was attempted. Initially, a set of standard accompaniment control settings were applied. These settings were identical to those used for the trials with recorded performances. The minimum time span for estimating tempo was set at 1 second and the preferred time span at 3 seconds. The two absolute position difference thresholds for determining corrective performance actions were set at 50 ms and 750 ms. Corrective tempo adjustments were configured to increase or decrease the accompaniment tempo by 10% relative to the estimated tempo of the singer. The initial settings invariably were unacceptable for live performance. Typically, singers would slow the tempo or extend a note at some point in the piece, causing the accompaniment system to slow also. Unlike the recordings, however, the singer would perceive hesitation in the accompaniment and respond by either decreasing tempo or hesitating. Over time, the tempo gradually decreased to an unacceptable level.

To address this problem, systematic adjustment of the accompaniment control settings were made over multiple performance trials. First, slowing of the tempo often occurred at the ends of phrases before a rest. The initial adjustment consisted of adding control markings to the performance score, indicating that the accompaniment should increase the tempo once the singer completed the phrase. When this adjustment proved insufficient, the minimum time span for estimating tempo was changed from 1 second to 1.5 seconds, making the system slightly less responsive to tempo changes overall. Subsequent adjustments involved raising this minimum time in increments of 0.5 seconds up to a maximum of 2.5 seconds. Finally, the last change was to adjust both the lowest score distance threshold and the percent by which the tempo is increased or decreased in order to resynchronize with the singer. The threshold was changed from 50 ms to 80 ms, and compensatory tempo changes were restricted to ±5% of the estimated tempo. Adjusting of the accompaniment control settings terminated once the singer indicated that there had been no unacceptable slowing of the tempo. During the adjustment process, most performers completed between 3 to 5 performance trials, including the initial trial. One performer completed a total of 7. During this sequence of performances, the tracking system configuration was alternated systematically between processing all observation types and observing only pitch. This prevented configuring accompaniment control in a way that was optimal for one specific version of the tracking system.

Another reason for multiple performance trials relates to the level settings and the amplitude threshold. In some cases, the initial settings of the levels and threshold were not acceptable during actual performances. Upper and lower ranges for the sound signals were louder or softer during a performance than during the initial performance of excerpts. There are two possible causes of these discrepancies. No accompaniment was generated while the initial settings were determined. In some cases, the additional sound during actual performances might have increased the level of signal detected by the microphone.

Alternatively, singers may adjust their dynamic level in response to presence of the accompaniment. In addition to simply increasing their own level when accompaniment is present, they may not be able to replicate the extremes of their expressive dynamic levels when no accompaniment is present. Regardless of the cause, however, it appears that sound levels and thresholds must be set within the context of an actual performance.

Finally, each singer gave 3 performances that were recorded. Time stamped position estimates and performance traces also were generated. Three different sets of observation types were used—pitch alone, pitch and note onsets, and all three observation types. Ordering of the trials was randomized, and different sets of observation types were used in the orders shown in Table 7-7. Upon completion of the performances, each singer was asked two questions. First, whether any performance seemed noticeably worse with respect to the accompaniment, and second, whether any performance seemed noticeably better with respect to the accompaniment. Subsequently, the singers were informed that different combinations of signal processing had been applied during the three performances. During all trials, the system accompanied the singer to the end of the performance. Most trials were reasonable performances, though some trials with the sixth performer contained noticeable synchronization errors (as will be shown). One trial with performer three contained a noticeable synchronization error that caused the performer to stop. The performer was asked to continue, and the accompaniment system resumed the performance, quickly recovering the correct position and tempo. This trial used observations of pitch and onsets, and the problem was later traced to a spurious onset reported in the middle of a note.

Handmade segmentations of the recordings were prepared for the trials using only pitch and the trials using all observation types. Time differences between the time stamped position estimates and the

Table 7-7. Ordering of the three trials for each performer when evaluating the system using live performances.

	Pitch	Pitch and Onsets	All Observations
Performer 1	2	1	3
Performer 2	3	1	2
Performer 3	2	3	1
Performer 4	1	3	2
Performer 5	3	2	1
Performer 6	2	3	1

handmade segmentations were generated for each trial. Graphs and summary statistics over the time difference data were subsequently assessed and compared. The results are similar to those for the recorded performance tests. For the trials using all observation types, histograms of the time differences exhibit a sharper distribution around zero, but sometimes contain more extreme outliers. Graphs of time differences against score position over a single performance typically do not show clusters of extreme outliers. The outliers are dispersed throughout the piece. Summary statistics of time differences for the six performers are presented in Table 7-8. The standard deviations were smaller for 4 out of 6 performers when using all observation types. Discarding the 5% outliers, the standard deviations were smaller for 5 out of 6 performers when using all observation types. A paired comparisons sign test (a test considering how frequently 5 out of 6 trials would show improvement if the two systems were identical) gives a P-value just under 0.11 for this situation. When all data is considered, the average standard deviation was larger when using all observation types than when using only pitch. However, the average is affected by the large standard deviation for performance six. This large value is due to a few extreme outliers. Some

Table 7-8. Mean and standard deviation of tracking time differences for live performances. Outliers were removed based on absolute time difference.

	All Data		5% Outliers Removed	
	Pitch	All Observations	Pitch	All Observations
Performer 1				
Mean	-42 ms	9 ms	-27 ms	26 ms
SD	102 ms	97 ms	60 ms	51 ms
Performer 2				
Mean	-132 ms	-89 ms	-110 ms	-59 ms
SD	181 ms	154 ms	111 ms	94 ms
Performer 3				
Mean	17 ms	14 ms	28 ms	26 ms
SD	90 ms	79 ms	71 ms	50 ms
Performer 4				
Mean	-34 ms	-13 ms	-21 ms	3 ms
SD	104 ms	80 ms	62 ms	48 ms
Performer 5				
Mean	-19 ms	29 ms	-14 ms	41 ms
SD	113 ms	116 ms	77 ms	69 ms
Performer 6				
Mean	-24 ms	-93 ms	8 ms	-26 ms
SD	179 ms	318 ms	116 ms	133 ms
Average				
Mean	-39 ms	-24 ms	-22 ms	2 ms
SD	128 ms	141 ms	83 ms	74 ms

of these outliers occurred because of an amplitude threshold that was too high, causing spurious reporting of note onsets in the middle of notes. Others were due to *portamento* in the performance where the actual fundamental pitch sung shifted by more than a third from the scored pitch.

Paired comparisons tests were run based on pairing position estimates for the same notes in the score, using the two trials for each performer. T-tests were run for each set of differences of differences based on a null hypothesis that the mean is zero. Results of these tests are presented in Table 7-9. Three of six tests showed improved performance using all observations with a P-value less than 0.15. Only one test showed a degradation in tracking accuracy when using all observation types—performance six. The average of the mean differences over the pairs of trials was 1 ms. Excluding performance six, this average was -11 ms.

The average means and standard deviations of time differences are smaller for the live performances than for the recorded performances. However, given the small number of performances and the large variability among performances, the differences cannot be interpreted too conclusively. Differences between tracking of live and recorded performances may exist, however, with two possible explanations readily proposed. First, the adjustments to tempo estimation prior to the live performances may have provided better tracking overall. Second, the live performers certainly listen to and respond to the accompaniment, unlike the recordings. To the extent that they try to synchronize with the accompaniment, their actions may result in improved tracking accuracy, as determined by the metric applied in this study. Also, the preliminary performances may have provided an opportunity for the singers to adjust to the accompaniment system.

Graphs of synchronization time differences for the live performances are similar to those for the recorded performances. These graphs are not as sharp as the graphs of tracking differences and contain

Table 7-9. Results of paired comparisons tests when pairing estimated positions for the same note across two trials. Each test considered two trials of the same performance, one trial using only pitch and one using all observation types. Three tests were run for each of six live performances.

P-Value per Paired Comparisons Test	Tests Showing Improvement Using All Observations	Tests Showing Degradation Using All Observations
<0.01	1	1
<0.05	1	1
<0.10	2	1
<0.15	3	1

fewer and less extreme outliers. These results are consistent with the claim that the accompaniment control system introduces a hysteresis relative to position estimation. Summary statistics for the synchronization time differences are presented in Table 7-10. Note that the standard deviations for individual performers, as well as on average, are comparable to or better than the respective standard deviations for the tracking time differences. Considering all data when using all observation types, the standard deviations were actually smaller for the synchronization differences than for the tracking differences in 5 out of 6 instances. Also, synchronization improved in 4 out of 6 instances when using all observation types as compared to using only pitch, but this is not terribly significant.

The improved agreement between tracking and synchronization results for live performances, compared to the same results for recorded performances, possibly can be attributed to the preliminary adjustment of accompaniment control settings and the ability of live performers to respond to the accompaniment system. The former situation would imply that it is important to understand the specific

Table 7-10. Mean and standard deviation of synchronization time differences for live performances. Outliers were removed based on absolute time differences.

	All Data		5% Outliers Removed	
	Pitch	All Observations	Pitch	All Observations
Performer 1				
Mean	-103 ms	-53 ms	-94 ms	-42 ms
SD	121 ms	92 ms	85 ms	81 ms
Performer 2				
Mean	-153 ms	-91 ms	-136 ms	-79 ms
SD	121 ms	103 ms	98 ms	92 ms
Performer 3				
Mean	-42 ms	12 ms	-39 ms	-5 ms
SD	95 ms	129 ms	77 ms	89 ms
Performer 4				
Mean	-55 ms	-36 ms	-81 ms	-31 ms
SD	134 ms	69 ms	81 ms	61 ms
Performer 5				
Mean	-12 ms	31 ms	-11 ms	32 ms
SD	89 ms	82 ms	74 ms	63 ms
Performer 6				
Mean	-88 ms	-93 ms	-78 ms	-102 ms
SD	98 ms	261 ms	90 ms	100 ms
Average				
Mean	-76 ms	-49 ms	-73 ms	-38 ms
SD	110 ms	123 ms	84 ms	81 ms

requirements for adequate accompaniment performance, both to produce reasonable accompaniment and to take advantage of more accurate score position estimation. The second situation would imply that, as previously argued, it is not sufficient to rely only on recorded performances for evaluating either accompaniment control or score-following systems. The two cases are distinct enough that techniques necessary or helpful in one case do not improve tracking accuracy or accompaniment control in the other.

The trials based on observing both pitch and onsets were not examined, primarily because of the one performance where the performer stopped in the middle of the piece. This performance clearly would dominate the statistics, as have other extreme cases. Qualitatively, these trials overall were similar to the trials based on other sets of observations. Performers did not show a consistent preference for any particular combination of observation types. Table 7-11 summarizes the performers' assessments of best and worst performances within the three trials. Performers 2 and 5 indicated that all performances were comparable. Performer 3 (the performer who stopped during one performance) indicated that the performance where the system observed both pitch and onsets was worst and the other two performances were comparable. Performer 1 indicated the performance where the system observed both pitch and onsets was best while the performance where the system used all observations was worst, stating voluntarily that this choice was because the system seemed to follow too closely. Performance preferences do not appear correlated with ordering of the performances within the trials. Lack of any clear preference of performances is not surprising, given both the simple approach used to control the accompaniment performance and the complex factors influencing assessment of musical performances.

As a final assessment, the accompaniment system was used to accompany selected recordings of the live performances. During these trials, the settings for the accompaniment control system were

Table 7-11. Performers' subjective assessments of best and worst accompaniment performances during live performance trials. Dashes indicate the performer expressed no preference. Numbers in parentheses indicate the order of the selected performance amongst the three trials.

	Best Performance	Worst Performance
Performer 1	Pitch and Onset (1)	All Observation Types (3)
Performer 2	———	———
Performer 3	———	Pitch and Onset (2)
Performer 4	All Observation Types (2)	———
Performer 5	———	———
Performer 6	Pitch and Onset (3)	———

identical to the settings used during the original performances. The sound level and amplitude threshold settings were determined by the same process previously applied when accompanying recorded performances. Two recordings of each performer were accompanied—the performance initially tracked by observing only pitch and the performance initially tracked using all observations. Trials using only pitch and trials using all observation types were run for each recording, producing a total of 4 sets of trials for each performer. Standard deviations of time differences for these trials are presented in Table 7-12. Note that the reported values for each recording are the averages over three trials. Values in parentheses indicate results of the original live testing under equivalent circumstances. In many instances, results of accompanying the recordings are comparable to the results obtained from the original live trials. The averages across performers are within 10 ms of the original averages. For performances originally tracked using only pitch, the new trials using all observation types yielded smaller standard deviations compared to the new trials using only pitch, regardless of whether or not outliers were removed from the time differences. For performances originally tracked using all observation types, the new trials using all observation types yielded smaller standard deviations in 5 out of 6 instances with outliers discarded. However, when all time differences are considered, using all observation types yielded smaller standard deviations in only 2 cases.

Tracking accuracy can also be compared across different recordings by the same performer when using the same observation types for tracking. Generally, the tracking accuracy for the recordings initially tracked using only pitch was comparable to the tracking accuracy for the recordings initially tracked using all observation types. The exception is the new trials using all observation types, when considering all time differences. Several extreme outliers occurred in the trials of the performances originally tracked using all observation types, yielding larger standard deviations relative to trials of performances originally tracked using only pitch.

In summary, accompaniment trials involving live performances exhibited several similarities to the trials involving recordings of singers performing with live accompanists. First, distributions of the time differences show similar behavior as more observation types are used for tracking. Specifically, the center of the distribution sharpens, but additional or more extreme outliers appear for some performances. The outliers result from the same problems that caused outliers in the trails with recorded performances. Second, note by note comparisons of trials using pitch and trials using all observation types indicate that using all the observation types seldom degrades overall tracking accuracy during a performance, providing tracking that is either comparable or improved. Third, distributions of the synchronization time differences are less sharp than corresponding distributions of the tracking time differences, suggesting a hysteresis is introduced by the accompaniment system. Overall, use of multiple observation types to improve tracking accuracy did not translate to improved synchronization, presumably due to the hysteresis

Table 7-12. Standard deviation of tracking time differences for recordings of live performances. Results are given for two recordings of each performer—the performance initially tracked by observing only pitch and the performance initially tracked using all observations. Standard deviations for each performance are the average over three trials. Values in parentheses indicate tracking results during the original live performances. Outliers were removed based on absolute time differences.

Recording	All Data		5% Outliers Removed	
(by original observation types used and performer)	Pitch	All Observations	Pitch	All Observations
Fundamental Pitch:				
Performer 1	105 ms (102 ms)	104 ms	62 ms (60 ms)	51 ms
Performer 2	177 ms (181 ms)	160 ms	137 ms (111 ms)	90 ms
Performer 3	115 ms (90 ms)	99 ms	92 ms (71 ms)	67 ms
Performer 4	82 ms (104 ms)	73 ms	66 ms (62 ms)	49 ms
Performer 5	110 ms (112 ms)	107 ms	78 ms (77 ms)	63 ms
Performer 6	175 ms (179 ms)	171 ms	107 ms (116 ms)	80 ms
Average	**127 ms** (128 ms)	**119 ms**	**91 ms** (83 ms)	**67 ms**
All Observations:				
Performer 1	94 ms	100 ms (97 ms)	68 ms	51 ms (51 ms)
Performer 2	181 ms	196 ms (154 ms)	135 ms	117 ms (94 ms)
Performer 3	127 ms	80 ms (79 ms)	92 ms	57 ms (50 ms)
Performer 4	92 ms	85 ms (80 ms)	72 ms	52 ms (48 ms)
Performer 5	118 ms	153 ms (116 ms)	82 ms	56 ms (69 ms)
Performer 6	176 ms	289 ms (318 ms)	92 ms	103 ms (133 ms)
Average	**131 ms**	**151 ms** (141 ms)	**90 ms**	**72 ms** (74 ms)

introduced by the simple accompaniment control mechanism. Subjective assessments provided by the few live performers do not conclusively support any claims of differences in accompaniment performance depending upon which observation types are used.

Several differences between the two sets of trials also are apparent. The average tracking accuracy improved slightly (though perhaps not significantly) for the live performances, even when the system accompanied recordings of the live performances. Also, the synchronization time differences are closer to the tracking time differences in the case of the live performances. These distinctions might result from the initial tuning of accompaniment control parameters (including the settings for tempo estimation) that occurred prior to the live performance trials. Also, the live performers may respond to the accompaniment system in a way that improves synchronization and tracking. If the accompaniment system is fairly consistent, these adjustments may influence tracking and synchronization accuracy even when subsequently evaluating trials that use recordings of those live performances. Alternatively, differences amongst the scores and performers may account for the improvement. For instance, no scores used in live performances contained recitativos or cadenzas.

7.6 Comparison of Expected and Actual Tracking Ability

Evaluations presented in previous sections have shown that tracking accuracy often improves when using all observation types. As already mentioned, ideally one would like to estimate improvements in tracking accuracy prior to actual testing. Testing either reduces the examples available for density estimation or requires live performers and is time consuming. In addition, the system's ability to track live performance must be evaluated eventually. In the previous section, differences in tracking accuracy were noted when following live performances versus recordings of singers performing with human accompanists. These differences may be due to characteristics of the two performance scenarios. In Section 6.4, the expected likelihood of confusing the performer's actual score position with another position was defined. This value was proposed as a good predictor of improvements in the actual average tracking accuracy. This value is computed based on the estimated observation distributions and example scores (or possibly comprehensive statistics for the conditioning variables). In this section, the correlation between the expected likelihood of confusion and the tracking accuracy actually observed is examined.

The expected likelihood of confusing the performer's actual score position with another score position is estimated by the following equation:

$$E[\kappa] = \frac{1}{\|Score\|^2} \int_{i=0}^{\|Score\|} \int_{j=0}^{\|Score\|} \kappa(i, j, f_{V|I}) \cdot [j \neq i] \, \partial j \, \partial i$$

The function κ defines the likelihood of assigning equal or higher density to point j than to point i based on a single observation. As presented in Chapter 6, this function is defined in terms of the observation distributions, which themselves are conditioned on information provided in the score. In order to actually apply this metric to predict improvements in tracking accuracy, two modifications were required. First, for the stochastic score-tracking models, numerical methods based on sampling are used to approximate the integrals. Second, the calculation is limited to considering points over a fixed size window, as used by the tracking system, rather than over the entire score. These modifications produce the following estimate of the expected likelihood of confusion:

$$E[\kappa] = \frac{1}{\|Score\| \cdot \|Window\|} \sum_{i=0}^{\|Score\|} \sum_{j=i-\|Window\|/2}^{\|Window\|/2} \kappa(i, j, f_{V|I}) \cdot [j \neq i]$$

Note that any points beyond the boundaries of the vocalist's score are treated as rests. This modified form of the expected likelihood of confusion is the predictor value actually considered in the subsequent analysis.

The expected likelihood of confusion was calculated for each score from the performance trials. The value was calculated multiple times for each score, once for each set of observation types used in the trials. These values can be compared against the actual standard deviations of the tracking time differences. A simple comparison would be to calculate the correlation between the two parameters. However, the correlation assesses the fit for a simple linear model relating the two values:

$$\sigma = a \cdot E[\kappa] + b$$

The observed standard deviations probably do not relate to the expected likelihood of confusion in this manner. A slightly modified model may be more appropriate. Recall that the expected likelihood of confusion indicates the average likelihood that the performer's actual score position will be assigned lower density than another position, based on a single observation. Let the variable d represent the expected distance between the performer's actual position and the estimated position when the system does not estimate the correct position. In this case the following model may be appropriate

$$\sigma^2 = d^2 \cdot E[\kappa] + 0 \cdot (1 - E[\kappa]) = d^2 \cdot E[\kappa]$$

The variance of course is the expected squared deviation from the mean, so this model assumes that the expected position estimate is the performer's actual position. Also, the model assumes independence between the likelihood of confusion and the expected distance between actual and estimated positions.

Correlation and regression calculations are based upon an assumption that residual errors are independent and normally distributed. However, errors in the standard deviations have been shown to be extreme in the positive direction, due to aspects of the musical performances that are not adequately modeled. Consequently, when assessing correlation and fitting lines, the logarithm of the standard deviations may better account for the variance in the data. Thus, in addition to examining correlation for the following model:

$$\sigma = d \cdot \sqrt{E[\kappa]} + c$$

where c represents any bias due to hand segmentation or time alignment, correlation was examined also for the following model:

$$\ln \sigma = a \cdot \sqrt{E[\kappa]} + \ln b$$

This model relates the expected likelihood of confusion to the logarithm of the actual standard deviations. Note that rewriting this model in terms of σ by removing the logarithm yields a model that is exponential in the expected likelihood of confusion:

$$\sigma = be^{a \cdot E[\kappa]}$$

Graphs of the fitted function are presented as exponential curves relating the expected likelihood of confusion to the actual standard deviations.

Correlations for both types of models are presented in Table 7-13. All models show positive correlation between the actual standard deviations and expected likelihood of confusion for all sets of trials. However, the correlations are stronger for the exponential model than the linear model, as expected. Correlations are stronger when the 5% outliers are discarded. Correlation is weaker for the live performance trials than for the trials using recordings. The low values are due to the trial using all observation types with performer 6. When this trial was removed from the set, all calculated correlations were above 0.7. The square of the correlation indicates the percent of the variance in the actual standard deviations that is accounted for by the model. In the case of recorded performance trials with outliers discarded, the exponential model accounts for over 59% of the variance; with outliers included the

Table 7-13. Correlations for standard deviation and expected likelihood of confusion assuming two different models. Values are given for trials with 8 recorded performances over 4 sets of observation types (trials using spectral envelope only were omitted), trials with 6 live performances over 2 sets of observation types, and the set of combined trials.

	Recorded Performances (32 standard deviations)		Live Performances (12 standard deviations)		All Performances (44 standard deviations)	
Model	All Data	No Outliers	All Data	No Outliers	All Data	No Outliers
Linear	0.636	0.744	0.304	0.59	0.547	0.687
Exponential	0.676	0.771	0.426	0.639	0.603	0.726

exponential model accounts for over 45% of the variance. Deficiencies in the stochastic score-following model likely account for some of the remaining variation, but the simple predictive models considered certainly do not completely characterize the behavior of the stochastic score-following model

For instance, Table 7-14 presents correlation values for sets of the trials grouped by the observation types used. Note that the correlations are higher for these sets than for the combined trials, except in the case of using only spectral envelope. Although the smaller number of values per set may produce such an effect, it is probable that some of the constants in the fitted models vary according to each combination of observation types. In particular, the expected distance between actual and estimated score positions may change depending upon the observation types. Thus, combining all trials yields a less effective predictor. Though the number of trials is small for each set of observation types, the high correlations suggest that the proposed models may be effective predictors of the individual scores that will hinder performance tracking based upon a particular set of observation types, relative to other scores. Also note that, for the recorded performance trials, the mean standard deviation with outliers removed decreases as the expected likelihood of confusion decreases. It was previously suggested that the likelihood of confusion may be a better predictor of the average tracking accuracy than the tracking accuracy for individual scores.

For the model based on observing only spectral envelope, the mean value of $E[\kappa]$ over the scores approaches 0.5. For specific scores, the value of $E[\kappa]$ sometimes exceeds 0.5, indicating that on average more than half of the score positions within the window will be confused with the performer's actual

Table 7-14. Correlations for standard deviation and expected likelihood of confusion assuming two different models. Values are given for trials with 8 recorded performances using 5 sets of observation types and trials with 6 live performances using 2 sets of observation types, 5% outliers removed.

	Recorded Performances (8 standard deviations)					Live Performances (6 standard deviations)	
	Pitch	Spectral Envelope	Pitch & Envelope	Pitch & Onsets	All Obs.	Pitch	All Obs.
Mean $E[\kappa]$	0.348	0.486	0.251	0.264	0.197	0.318	0.187
Mean SD (no outliers)	117 ms	767 ms	103 ms	103 ms	80 ms	83 ms	74 ms
Corr. Linear Model	0.79	0.337	0.871	0.801	0.874	0.985	0.685
Corr. Exp. Model	0.848	0.436	0.91	0.828	0.863	0.972	0.699

position. The value of 0.5 might be considered a critical point for E[κ], indicating that the observation types are ineffective at distinguishing the performer's actual score position and are almost useless. The observations probably will not correct most tempo estimation errors that cause the convolution result to estimate the wrong position. Since the performer is likely to change tempo noticeably at some point during a performance, serious position estimation errors and synchronization errors should be expected. Many of the performance trials using only spectral envelope produced noticeable or sometimes extended periods where the recording and accompaniment system were not synchronized. Note that when observing only pitch, the mean value of E[κ] is closer to one-third than one-half. Trials using only pitch were significantly better than trials using only spectral envelope.

Since the fitted exponential model produced correlations comparable to or superior to correlations for the linear model, graphs of the fitted exponential curves were examined. Figure 7-7 shows the curves fit to the trials using the recordings. Each point on the graph represents the results of tracking one performance using one set of observation types. Note the positive correlation between E[κ] and the actual standard deviation, and the tighter clustering of points around the curve when the time difference outliers are removed for each trial. Figure 7-8 shows curves fit to the trials on live performances (the curves labeled "b") and Figure 7-9 shows curves fit to the combined trials over both recorded and live performances (the curves labeled "b"). The live performance trials produce curves that increase less rapidly but have a higher intercept, appearing relatively linear. This effect may be due to the smaller average values of E[κ] for this set of scores, the altered settings for tempo estimation used during live performance trials, or the ability of live singers to respond to the accompaniment system. However, these few trials do not deviate from the original curves significantly enough to produce drastically different curves when the exponential model is fit to the combined sample. Note that including the trials using only spectral envelope would yield curves that increase more sharply.

The standard deviation appears to decrease slowly as the value of E[κ] decreases. A positive intercept is reasonable considering the existence of measurement errors due to time alignment and hand segmentation. Several important conclusions can be drawn if the estimated curves are accepted as reasonable indicators of the change in standard deviation relative to the change in E[κ]. Increasing the number of observation types produces similar changes in the average value of E[κ] for both sets of scores, but this change in value spans less than one-quarter of the distance between the origin and the highest E[κ] value for any trial. From this perspective, the observation types actually used leave us fairly far out on the curve. However, when the 5% outliers are removed from the time differences, the fitted curve passes below 25 ms well before the origin, and the mean E[κ] when using all observation types is already below the 100 ms mark. The expected standard deviations for both score following and performance synchronization by humans are probably within this range. Consequently, the two objectives of future enhancements to the score-following system should be reducing the number of extreme outliers in the tracking time differences and reducing the mean value of E[κ] as calculated over vocal scores.

Fortunately, enhancements that reduce the mean value of E[κ] may also reduce the number of extreme outliers. In the graphs, the variance of the actual standard deviations about the curve appears to

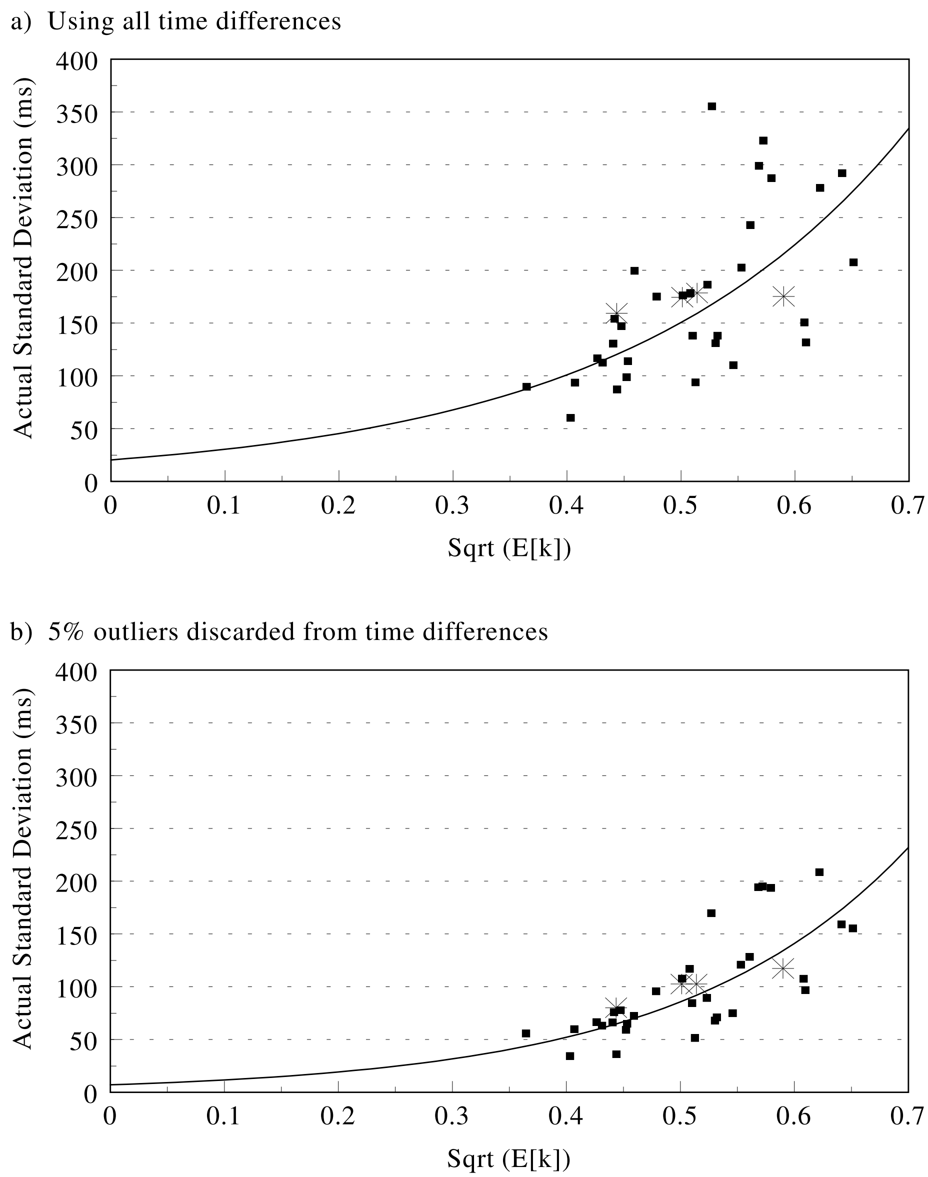
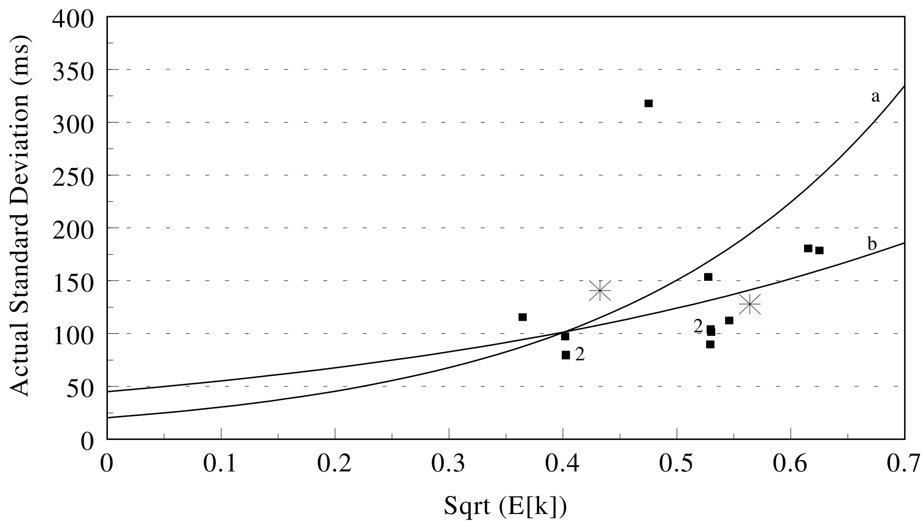


Figure 7-7. Exponential regression curves based on expected likelihood of confusion mapped to actual standard deviations of time differences from tracking the eight recorded performances. Each point corresponds to tracking one performance using one set of observation types. Asterisks indicate averages for trials using different observation types (from left to right: all types, pitch and spectral envelope, pitch and onsets, and only pitch).

decrease as E[κ] decreases. Generally, as the observations better distinguish the performer's actual score position, other sources of error such as poor tempo estimates are less likely to influence position estimation. As previously discussed, position estimation can be improved by adding new observation types, adding conditioning variables to the existing observation distributions, and improving density function estimates by using better analytical models and additional data. Effective application of these

a) Using all time differences



b) 5% outliers discarded from time differences

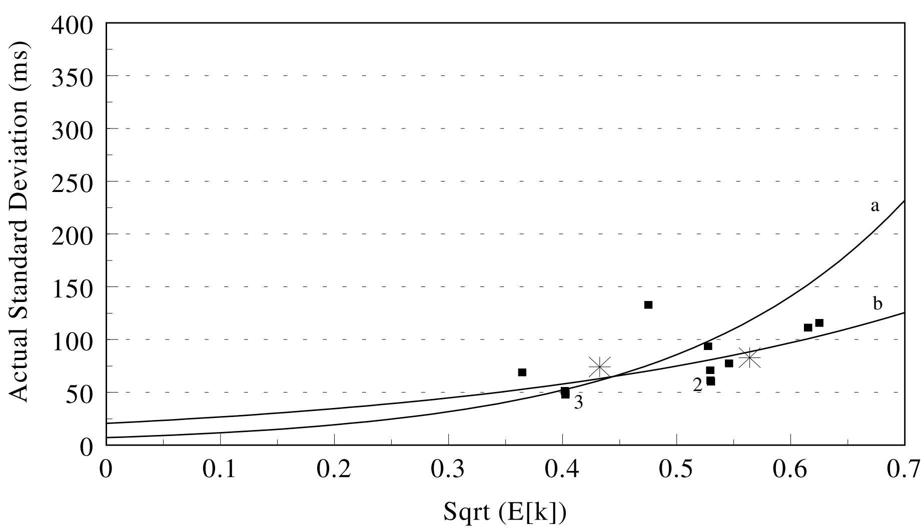
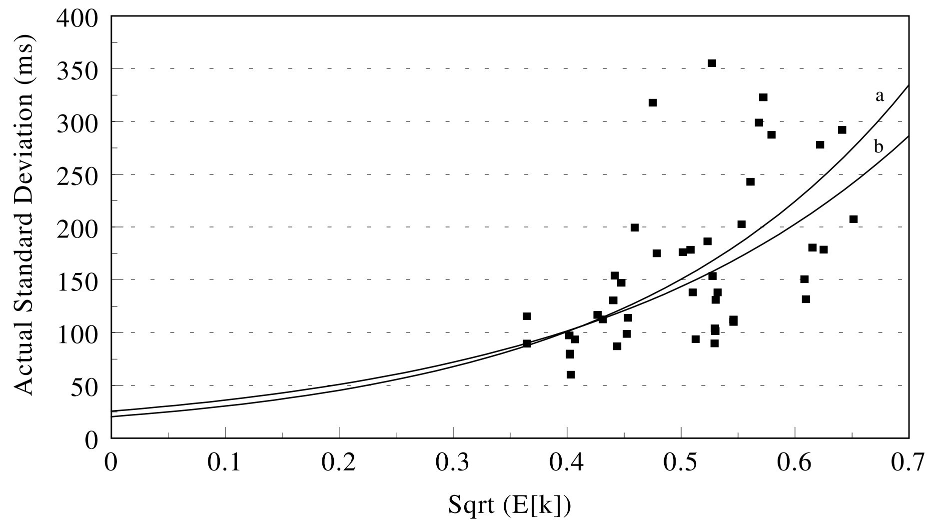


Figure 7-8. Exponential regression curves based on expected likelihood of confusion mapped to actual standard deviations of time differences from tracking the six live performances. The curves labeled "a" are curves previously fit to the recorded performance trials (taken from Figure 7-7); the curves labeled "b" are curves fit to the live performance trials. Each point corresponds to tracking one live performance using one set of observation types. Numerals indicate overlapping points. Asterisks indicate averages for trials using different observation types (from left to right: all types and only pitch).

techniques will reduce the mean value of E[κ] as well. Furthermore, examination of outliers in the trials using recordings showed that the most significant causes of outliers included inappropriate amplitude thresholds, errors in tempo estimation or expected durations, and observation distributions that did not characterize the actual observations made during a particular event. Improving observation distributions certainly will improve E[κ]. Adjusting or eliminating the use of amplitude thresholds would require new

a) Using all time differences



b) 5% outliers discarded from time differences

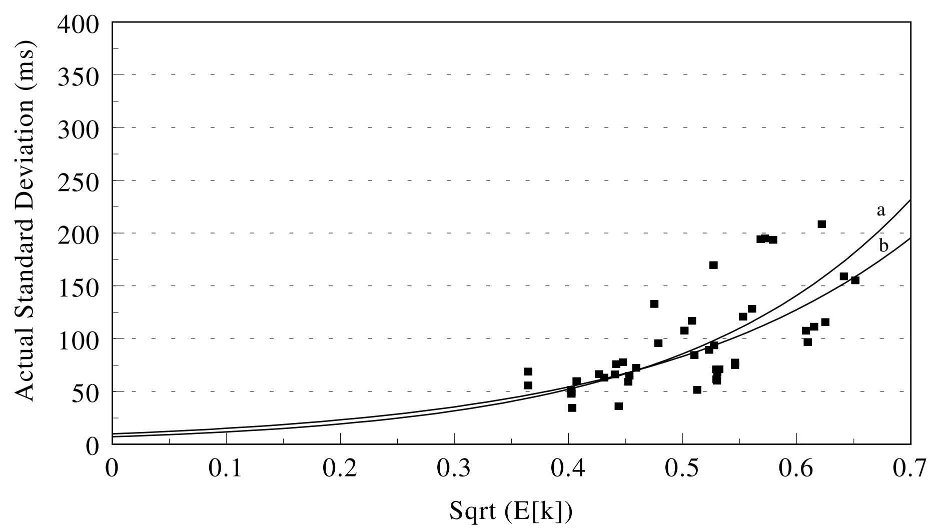


Figure 7-9. Exponential regression curves based on expected likelihood of confusion mapped to actual standard deviations of time differences from tracking the live and the recorded performances. The curves labeled "a" are the curves previously fit to only the recorded performance trials (taken from Figure 7-7) while the curves labeled "b" are the curves fit to the combined set of trials over the live and recorded performances. Each point corresponds to tracking one performance using one set of observation types.

estimates of the observation distributions and possibly entirely new definitions for some observation types, hopefully improving E[κ].

7.7 Summary and Conclusions

An automated accompaniment system incorporating the stochastic score-following model has been presented. Pertinent details of the design and implementation of this system were provided, including methods for estimating a performer's score position and tempo, and a basic form of accompaniment control. Methods of evaluating the system's tracking accuracy and performance synchronization were discussed. These techniques were applied to performance trials involving both recordings and live performances by singers. Subsequently, values for the expected likelihood of confusing two score positions were calculated for each score appearing in the trials. These calculations were shown to be positively correlated with the actual tracking accuracy measured during the performance trials.

Several conclusions can be drawn from the provided analysis. Using multiple observation types improves tracking accuracy if the observations discriminate score position and their density functions are accurately defined and estimated. More precisely, average tracking accuracy improves as the expected likelihood of confusing two score positions decreases. Proper definition and modeling of multiple observation types reduces this value. Improved tracking does not guarantee improved synchronization, especially if the accompaniment control system is simple and cannot take advantage of the improvement, or the control settings are inappropriately configured. Also, outliers may cause problems in specific cases even if the average tracking accuracy improves, possibly resulting in a perceived degradation in accompaniment performance for those instances. Primary causes of outliers include inappropriate amplitude thresholding, insertion of rests and breaths by the performer, incorrect tempo estimates or expected durations, and observation distributions that do not adequately characterize the actual observations made during the performance of an event. Live performances may be easier to track than are recorded performances. Improved tracking may result from the additional information provided by the live performers, such as specific values for expected tempi and durations that are not indicated explicitly in the score. Alternatively, live performers probably adjust to the accompaniment, attempting to synchronize with the computer and not to violate their own beliefs about the system's expectations. Improved tracking accuracy may be a secondary effect of these adjustments.

Although the implemented score-following model is probably not as competent at tracking vocal performances as a human, the difference in tracking ability may not be too great. Examination of outliers in the time differences and the analysis based on the expected likelihood of confusion helped to identify enhancements to the tracking model that possibly will improve tracking accuracy. First, the distance density model could be specialized for the cases of recitativos and cadenzas, and possibly also for gradual tempo changes like *ritardandos*. The latter situation might include developing models of expected tempo changes based on analyzing the form of a musical score (*e.g.*, automatically identifying the ends of phrases and the location of important cadences). Second, one could analyze a larger corpus of recorded performances (possibly all recordings used throughout the this study) and use additional conditioning variables to improve density functions estimated for the currently defined observation types. Third, new observation types could be developed based on changes to the signal processing, such as using longer term pitch or state changes indicated by autocorrelation or spectral difference. Following such extension of the stochastic score-following model, the evaluation process would include calculation of expected likelihood of confusion over the scores, in order to assess which combinations of techniques are expected to provide the most significant improvements. Subsequently, another round of live testing could be used to compare tracking systems based on observing fundamental pitch, all three observation types defined in this study, and the newly extended or modified set of observation types.

Chapter 8

Related Techniques and Conclusions

8.1 Summary of the Stochastic Approach to Score Following

Automated accompaniment systems accept a symbolic score of a musical composition, monitor performances by live soloists, and perform the other parts of the score in real time and in a musical manner. Part of the accompaniment task is identifying the score position and tempo of the soloists. These systems need to track the score position at a resolution finer than a note in the musical score. Thus score position tracking requires high precision in addition to high accuracy and low latency. For vocal performances, it is necessary to integrate multiple sources of information, including several different measurements of the performance sound signal. When combining multiple information sources to produce a single estimate of a performer's score position, each value or measurement should be considered relative to the other available measurements. Each value should influence the final estimate based upon its ability to reliably discriminate the score position. In addition, it is not possible to specify a deterministic function mapping the available information (both time-stamped signal measurements and the given score) to the performer's score position. Several factors can prevent specification of a true function, including difficulties in reliably measuring the performance signal, the limits of available computation, and an inability to fully specify the function because it is either very complex or highly arbitrary. Consequently, a stochastic approach has been used to track the score position of a vocal performer.

A general stochastic model for score following was presented. Several generalizations and assumptions are needed in order to define a model that is tractable to estimate and can be implemented efficiently. Component distributions of the model were defined and estimated, including a distribution of the actual amount of score performed and distributions for each of three observation types. The model combines the observations in a way that is statistically sound. The definition of observations and the estimation of distributions is done carefully and accurately, to the extent possible. Arguments were provided for why using multiple observation types (more variables) can improve tracking accuracy. A method for assessing the improvement was presented.

Evaluation of the stochastic performance tracking system demonstrated improved tracking when using multiple observation types. Namely, the results indicated a decrease in the difference between the time that a position was estimated and the time the singer actually performed that position. A decrease in statistics of these time differences over a single performance correlates with a reduction in the expected likelihood of confusing two positions based on the estimated observation distributions and the score. Thus, sharper observation distributions yield improved tracking. Although the simple accompaniment control system cannot always take advantage of this improved tracking, the developed accompaniment system was used to accompany live performances, often successfully.

The statistical model developed in this project is based on continuous probability. A corresponding discrete model was defined along with an efficient software implementation. Several different distribution functions were employed during definition and estimation of the distance and observation densities. In the next section, the stochastic score-following model is compared to several well-known statistical techniques and models. These approaches have been employed to estimate values in several different applications spanning a variety of objectives and disciplines.

8.2 Related Statistical Techniques

One approach to statistical modeling is to estimate models that are tailored for the domain or application. This approach is well-suited to the performance tracking problem. The requirements for high precision and accuracy of score position estimates make it necessary to include timing information and multiple measurements of the performance signal in the statistical models. Furthermore, constructing a general vocal performance tracking system is greatly facilitated by incremental development in the research phase. Use of appropriate notation, estimation processes, and techniques for analyzing distributions enables gradual development and comparison of alternatives. The limited data available and the desire to obtain distributions in the limit (*i.e.*, the distributions over all possible performances) require both simplifying assumptions and careful estimation of density functions. Also, this approach permits application of analytical techniques and specification of parametric density functions. Appropriate use of such techniques can sometimes compensate for limited data and may be more effective than estimating general representations of the distributions.

The reader who is familiar with statistical modeling and stochastic processes has probably observed similarities between the presented stochastic score-following model and other well-known statistical models and techniques. The methods for applying each of these models are similar, in the sense that a conditional distribution is defined as a solution to a specific problem and basic, commonly known statistical principles are applied for modeling and estimation. However, these methods do exhibit specific differences including the information (variables and distributions) included in the models, the assumptions necessary to support the models, the estimation process applied, and the implementation of the model (*i.e.*, method of calculating the solution). Selection and definition of a model can be based on specifics of the application, including properties of the necessary and available information (conditioning variables), assumptions that are valid for the application, and requirements on accuracy and latency for the implemented system (generated solution). This section provides a brief definition of several well-known statistical techniques and compares them to the stochastic score-following model. Specifically, consideration is given to Markov models (both discrete and continuous time versions), hidden Markov models, semi-Markov models, hidden semi-Markov models, Kalman filters, dynamic programming and Bayesian networks.

A *state-transition process* is a process that can be characterized as transitioning through a sequence of discrete states. At the highest level of abstraction, a *state* can be viewed simply as a characterization of the process at a specific instant in time. More specifically, a state often constitutes an assignment of values to one or more variables describing a physical system. Thus, a state may include values such as temperature, rate, or position. A change from one state to another in a state-transition process is called a *transition*. A probabilistic characterization of a state-transition process defines a set of conditional probabilities indicating the likelihood that each possible state will be the next state of the process conditioned on all preceding states of the process:

$$P\ [S_{t+1} = s_i \mid S_t = s_j, S_{t-1} = s_k, \ldots, S_1 = s_1],\ 1 \leq i, j, k \leq N$$

where N indicates the number of distinct, possible states of the process. Note that if all the conditional probabilities are arbitrary and either the number of distinct states is large (possibly countably infinite) or the state sequence for a process may be arbitrarily long, the statistical behavior of such a process will be complex and difficult if not impossible to define. The defined set of possible states in combination with the probabilities for all transitions and an initial state distribution specifies a *model* of the state-transition process.

A *Markov process* is a state-transition process for which the subsequent behavior of the process depends only upon the last state of the process. The behavior of the process of course includes the probability that each possible state will be the next state of the process:

$$P[S_{t+1} = s_i \mid S_t = s_j, S_{t-1} = s_k, \ldots, S_1 = s_1]\ =\ P[S_{t+1} = s_i \mid S_t = s_j]$$

Thus the current state (*i.e.*, its component variables) includes all context relevant to determining the probability of the next state of the process. Models of Markov processes (*Markov models*) can be much simpler than probabilistic characterizations of general state-transition processes. If a Markov model is used to determine properties for a process that is not known absolutely to be Markovian, or it is used to approximate a state-transition process that is not Markovian, then a *Markov assumption* is said to be made.

Since the Markov assumption is sometimes reasonable, and both the mathematics and computation of Markov models is often tractable, the investigation of Markov models has been quite extensive. Martin (1967) and Howard (1971) provide good overviews. In particular, analysis of the conditional distributions describing the transitions can provide answers to important questions about the targeted process. For instance, limiting behaviors of the process may be determined, such as whether the distribution over the possible process states converges over time to a fixed distribution. Markov models are commonly applied to problems in operations research, biology, and economics. Also, Markov models have been applied in a generative form for music composition in the twentieth-century, whereby probability distributions are defined to describe the sequencing of a set of musical elements (the variables in the states) and a piece is generated by simulating the process, randomly transitioning through a sequence of states.

Several variations and extensions of Markov models have been investigated. For instance, if the probability that a process will transition between two states remains dependent on the state but changes over time, a *time-varying Markov model* can be specified for the process. These models can be viewed as a composite of individual Markov models for different points in time, each component model specifying different probabilities of transitioning between states. As another example, transitions in a basic Markov model commonly are thought of as occurring once per every discrete time step, forming a *discrete time Markov model*. However, the time of the transitions can be allowed to vary, permitting for *continuous time Markov models* (Anderson [1991] and Howard [1971] provide overviews). Such models can be specified as containing a distribution for each of two values per transition—the next state of the process and the elapsed time between transitions. Note that in order for such models to remain true Markov models, the time between transitions (*i.e.*, the time that the process spends in each state) must depend exclusively on the current state. This restriction implies that the length of time the state has been occupied is independent of (and therefore does not influence) the successive state. Also, this duration does not influence the remaining amount of time that the current state will be occupied. The latter point implies that the time spent in each state must be exponentially distributed, although the exact exponential distribution may differ for each state.

Although Markov models exhibit an elegant simplicity that often enables tractable mathematical solutions to useful questions, it is excessively limiting to insist that all information relevant to predicting the behavior of the process must be characterized by the state. This requirement can lead to an unacceptable explosion in the number of states as variables are added, making the Markov model less appealing mathematically, computationally, and with respect to the specification of probabilities. Such an increase in the number of states is particularly frustrating if it results from duplicating the values of variables from other states within the current state of the process. Also, since the state space is discrete, reasonably accurate modeling of continuous process variables can require an excessive number of states. Restricting the allowable distributions of elapsed times between transitions can be limiting as well. When modeling processes that fall under any of these cases, a more general model may be preferred to a simple model that offers a poor approximation. For instance, it may be advantageous to consider models that can use arbitrary distributions to characterize the time between transitions and also can include the subsequent state of the process as a conditioning variable. These models represent time as continuous but are not truly Markovian. Consequently, they are referred to as *continuous time semi-Markov models*. Models that restrict the allowable times to integral values but still permit the subsequent state of the process to influence the elapsed time are called *discrete time semi-Markov models*.

Another well-known extension of Markov models was first presented by Baum and Petrie (1966). They considered modeling sequences of values (each value in the sequence drawn from a finite set of values) that unto themselves were clearly not Markovian. In other words, it was not acceptable to model the process with a Markov model indicating the next value in the sequence as a probabilistic function conditioned on the most recent value of the sequence, or even as a probabilistic function conditioned on all of the preceding values. One or more important, underlying variables influence the next value in the sequence, and these values are hidden from direct observation. Consequently, they proposed a model for a doubly-stochastic process containing an embedded Markov model whose states contain variables that are either hidden from observation or unknown. The values in the observed sequence are characterized by a probabilistic function conditioned solely on the state of the underlying Markov process at the time of each respective observation. Thus while the observable sequence of values is not assumed Markovian, the combination of the observed values and the hidden state variables is. Such models have since become known as *hidden Markov models*. Note that since each observed value is conditioned only on the current state, a hidden Markov model can be mapped uniquely into a Markov model by 1) constructing the states as a cross-product of the hidden Markov model states and the possible observed values and 2) defining the transition probabilities as a cross-product of the transition probabilities in the hidden Markov model and the probabilities of the observable values (conditioned on the states).

Hidden Markov models have been used for several applications in computer science. This popularity is largely due to a technique that Baum and his colleagues (1970) provided, enabling estimation of hidden Markov models from samples. They also showed that such estimation converges in probability to a locally optimal estimate, independent of the form of the true probability functions for the model.

Hidden Markov models also are appealing because the distinction between unknown state variables and observable values captures a salient aspect of many situations in real-world signal processing. The best known examples of applying hidden Markov models appear in the area of automated speech recognition. The task of interest in this case is identifying the sequence of spoken words given measurements of the speech sound signal and knowledge of the language. Commonly used state variables include the phoneme spoken, the phonetic context of the phoneme spoken, and the "sequence" within the phoneme spoken (note that this latter variable can be thought of as implicitly representing a variable duration portion of the spoken phoneme). Hidden Markov models have become the predominant approach to modeling both the acoustic and linguistic properties of speech (Baker 1975; Jelinek 1976; Bahl, Jelinek and Mercer 1983; Lee 1989).

More recently, extensions of hidden Markov models have been used for autonomous robot navigation (Simmons and Koenig 1995; Koenig and Simmons 1996). The task of interest is enabling a robot to identify its own location on a map of the environment as it moves through that environment. In the work by Simmons and Koenig, the state variables include physical location of the robot and directional orientation, and the observed values are calculated from sets of measurements by sensors such as sonar. Also, the probability that the robot has transitioned between specific states is further conditioned on the transition attempted by the robot (*i.e.*, where the robot tried to move). Since the attempted actions are selected by the robot (based upon a probability distribution over the states) and are known with certainty, this extension of a hidden Markov model is referred to as a *partially observable Markov decision process model*, or POMDP model. Recently, Thrun and colleagues have used similar models for estimation of maps in unknown environments (Thrun, Burgard and Fox 1998) and have applied techniques for selective use of sensor readings in position estimation (Fox *et al.* 1998).

Hidden Markov models can be extended to permit the use of arbitrary time distributions characteristic of the discrete and continuous time semi-Markov models. As the reader might anticipate by now, these models are referred to as *hidden semi-Markov models* and can be further subcategorized as either discrete time or continuous time depending upon their treatment of time. Recent work in computer science has applied such models to tasks involving segmentation of a sampled signal. The work by Chrisman (1996) developed techniques for these models and applied them to segment synthetic data generated by known models and also to segment sensor data from a flight of the Space Shuttle. The objective of the latter task is to identify important stages in the Shuttle's propulsion system. Changes between stages are indicated by significant inflection points (change of sign of the first derivative) in the sensor data stream. The amount of time spent in a given state is also an important cue in determining the transitions between process states. Hidden semi-Markov models have also been applied to automated speech recognition (Russell and Moore 1985; Levinson 1986).

Additional extensions of the discussed Markov models are scattered throughout the relevant literature. However, given the number of examples already considered, even the casual reader should have observed a pattern. Specifically, the basic Markov model is repeatedly extended to include additional information and permit greater flexibility for accurately modeling targeted processes (*i.e.*, to handle more variables and types of distribution functions). These extensions have been motivated in part by theoretical investigations, but also to a large extent by the desire to model more complex processes and to obtain accurate answers to a wider range of questions about the processes, preferably through efficient estimation and computation. Put less euphemistically, the limitations of preexisting statistical models have often become painfully obvious when applied to real-world tasks.

One noted shortcoming of the discussed Markov models is that the states are inherently discrete. Consequently, when using these models to characterize processes with continuous state variables, the variables must be discretized. Such discretization may require a large number of states in order to accurately model the targeted process. Models that require a large number of states (variables) complicate probability estimation and render straightforward calculations over the distributions computationally expensive or intractable, especially for real-time application. Computational techniques applied to the models generally are based upon either linear algebra and transform analysis, direct simulation, or graph algorithms (state-based models can be represented as graphs of course). The required computation time for all such methods increases as a function of the number of states. For extremely large numbers of states, such as might be needed for highly accurate approximation of continuous variables, the accuracy of the results can also degrade due to numerical errors. For instance, summations involving the discrete states and transitions generally correspond to integrations over the continuous variables. The preceding discussion of Markov models has not addressed the use of states that include continuous variables.

When applying the previously described models, probability estimation for large numbers of states is sometimes addressed by grouping states into equivalence classes and estimating distributions for a manageable number of classes. This technique is sometimes referred to as *tying* the states. Depending upon the variation among the distributions for each state, this technique may or may not yield an acceptably accurate model of the process when the number of states is limited so as to permit feasible computation.

An alternative approach is to work directly with continuous state variables and model the required distributions using continuous density functions represented in a parametric form. For processes where appropriate parametric functions can accurately model the true distributions, estimation and calculation over these functions may be both simpler and more accurate. Although such models can be described as examples of *continuous state Markov models*, a distinct line of work in the fields of control theory and signal processing has investigated a restricted form of such models that yields a solution method commonly referred to as a *Kalman filter*. Readers who might prefer a discussion based around the former term are encouraged to construct the definition of a time-varying, continuous time, continuous state, partially hidden semi-Markov decision process model as an entertaining exercise.

To use a basic Kalman filter, one must apply a linear system model to describe the targeted process. Specifically, this model defines the values of the state variables at a distinct point in time as a linear transformation of the state variables at the preceding point in time plus some noise. Observations reported by sensors at a distinct point in time are modeled as a linear transformation of the state variables at that time plus some noise. This model is defined by the following pair of equations:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{u}_t$$

$$\mathbf{v}_t = \mathbf{B}\mathbf{x}_t + \mathbf{w}_t$$

where $\mathbf{x}_t$ is the vector of state variables (the state vector) at time t, $\mathbf{v}_t$ is the vector of observation variables at time t, $\mathbf{A}$ is the state transition matrix, $\mathbf{B}$ is the observation matrix, and $\mathbf{u}_t$ and $\mathbf{w}_t$ are the additive noises at time t for the state transition and observation, respectively.

If the noises are uncorrelated through time and assumed to be Gaussian with a mean of zero, the optimal estimate (according to several criteria including minimum mean squared error) of the current state given a sequence of observations can be calculated in closed form. This calculation uses the matrices **A** and **B,** and also the means and covariances for each of the noise sources. The resulting estimate consists of a mean vector and a covariance matrix defining a Gaussian distribution over the state variables. Note that because the observations and the state transitions depend only on the current state (hence a Markov assumption is made), the estimate can be computed recursively over the sequence of observations. Details of the full computation are omitted here, but can be found in (Kalman 1960; Maybeck 1982; Cowan and Grant 1985; Mulgrew and Cowan 1988). However, since the equations to compute the estimate define a linear combination of the observations and the previous estimates, the computation can also be viewed as filtering the observation sequence with an infinite impulse response (IIR) filter that is optimal according to several criteria. Hence the term Kalman *filter* and the method's appeal in signal processing.

The model assumed by the Kalman filter is similar to a hidden Markov model but uses continuous state variables and includes a simple analytical model relating the variables. No underlying analytical model is implied by the hidden Markov model, which relates the variables only through probabilistic functions. Also, applications based on Kalman filtering typically use explicit definitions of the state variables, such as the coordinates for a position in two or three dimensions. The use of a linear system model and limited distribution functions provides additional information about the process, enabling a closed-form solution when estimating the state distribution. When these assumptions are appropriate, the state distribution can be computed much more efficiently by using a Kalman filter than by using standard simulation of a discrete time hidden Markov model with a large number of states. Accurate estimation of the component distribution functions may also be facilitated.

While the Kalman filter as described is based upon a discrete representation of time (hence is termed the *discrete time Kalman filter*), a continuous time version can be considered. However, the discrete time version is the most frequently examined filter since most applications of Kalman filtering deal with discrete time observations. Although the Kalman filter is based upon an assumption of Gaussian distributions, the method remains popular for engineering applications because it produces an optimal state estimate for processes involving linear systems in the presence of Gaussian noise. Also, the estimate is optimal with respect to the minimum error variance criterion even if the noise distributions are not Gaussian but the higher-order moments (moments other than the mean and covariance, such as skewness) of the distributions are either unknown or can be assumed insignificant for estimating the state. Common applications for Kalman filters include position tracking within navigation systems for ships and satellites, estimating process variables for automatically controlling mechanical and chemical processes in manufacturing, and filtering or equalizing noisy communications channels.

Various extensions of the basic discrete time Kalman filter have been examined. *Adaptive Kalman filters* use a model that permits the covariance matrices of the noise sources to vary over time and also calculate estimates for these matrices. *Extended Kalman filters* use a model of the process with some nonlinearity, such as expressing the observation as a nonlinear function of the state plus some noise. The coefficients of the optimal filter for the nonlinear models then require solutions to nonlinear differential equations that can only be approximated by closed-form solutions. Optionally, iterative numerical methods can be used to solve the equations for each time increment. Work in robotics has examined the use of Kalman filters within autonomous navigation systems, to estimate the position of the robot and the position of objects viewed by the robot (Smith and Cheeseman 1986; Crowley 1995). Kalman filters have been applied in machine vision to estimate motion of rigid objects in a sequence of images when the motion is restricted to movement in a plane parallel to the image plane (Legters Jr. and Young 1982). Extended Kalman filters have been applied to estimate motion of three-dimensional rigid objects from sequences of two-dimensional images (Broida and Chellappa, 1986). This motion was restricted to translation and rotation in a plane perpendicular to the image plane.

The different stochastic models used to describe real-world processes vary in their ability to accommodate different types of information, different distribution functions, and different analytical models. Selection of models for different applications must consider the efficiency of calculating targeted solutions given the model, feasibility of estimation, and precision and accuracy of the model with respect to the calculated estimate (the solution value or distribution). The importance of these tradeoffs is specific to the application. Score following requires position estimates that are produced with high precision, high accuracy and low latency. It benefits from incorporating continuous representations of rate and time that in turn require consideration of continuous valued score position. Since the precision of score position estimates must be finer than a single note anyway, and since relative note durations are specified in musical scores, it is helpful to consider a continuous model involving tempo, elapsed time, and score position. The real-time requirement on producing the position estimates demands efficient computation of these estimates. However, aesthetics of musical performance require that position estimates should be highly accurate. Ultimately this accuracy cannot be sacrificed for the sake of efficiency. As a further restriction, only limited performance examples can be collected for estimating distributions. Some distributions are associated with simple analytical models that can compensate for lack of data, but many are not. Finally, the necessity of using both arbitrary distributions and parametric distributions that are not well-behaved mathematically demands that solutions be calculated numerically rather than in a closed form.

The requirements for high precision, high accuracy, low latency score following are not entirely supported by the assumptions of the previously described models. For instance, exclusive use of models based upon discrete time or discrete state variables can be limiting. Although the implementation of the score-following model requires discretization of the continuous model, the associated continuous model is effective for defining the distribution of actual distance based on tempo and elapsed time. Furthermore, arbitrary variations in the elapsed time between observations are best accommodated by using a continuous representation of time. Forcing observations to be reported at consistent time intervals would require redefinition of the observation types, ultimately complicating density estimation and likely producing a wider variance among the observed values. For example, if the reports of estimated fundamental pitch were not staggered in time to accommodate recognized note onsets, some reported pitches would require processing of signal that either crossed onset boundaries or spanned less than 100 ms (the portion of signal preceding the onset having been discarded).

Correct timing is also very important in music. A continuous model is helpful in assessing the accuracy of the position density estimated by the corresponding discrete implementation. By first generating a continuous model (as accurately as possible) and then discretizing, familiar techniques of numerical analysis can be used to determine a discretization that limits approximation errors to an acceptable level. This approach is in contrast to trial and error experimentation using multiple implementations of the model based upon different discretizations. Also, a direct connection can be maintained between the solution calculated from the discrete implementation and the solution to the continuous model. The generated solution is an approximation to the density rather than a direct approximation to the probability within a region of the score. Although this approach may require a high sample rate, achieving equivalent accuracy and precision for position estimates would necessitate a similar number of states for any discrete model.

Consider an arbitrary probability function over the continuous variable indicating score position. Suppose that we attempt to represent the function as several contiguous states, each state encompassing 100 ms of score. Probabilities could be estimated for the states by calculating the area under the function within the region associated with each state. Note that nothing in this model retains information about the distribution of probability within the region corresponding to a state. Thus, no curve within a region can be distinguished from a constant function whose integral over the region has equivalent area. The overall function is thus *aliased* and cannot be reconstructed with complete accuracy from the state representation. Now if this model were the chosen representation for score position, then the most precise position estimate that one could obtain would be a single state. However, selection of the most likely state does not necessarily equate to the 100 ms region of score that is most likely to encompass the singer's actual position. If a neighboring state is almost as likely and the probability over the regions spanned by both states is skewed towards their common boundary, the actual most likely region may be shifted almost 50 ms from the region spanned by the most likely state. Although estimating the mean of the function through weighting the center of each region by the corresponding state's probability is less affected in this case, it is possible to construct cases where this calculation would also be less accurate.

The problem described can of course be addressed by decreasing the size of the region associated with each state, causing a corresponding increase in the number of states. The area in regions larger than the span of each state would be approximated by adding the areas for several of these smaller regions. In this respect, the utilization of discrete states is not inferior to direct sampling of the density function. The important point, however, is that a discrete state representation does not automatically eliminate calculation errors that can result from aliasing. Consequently, it is important to consider the continuous model in relation to any discrete approximation. Furthermore, the existence of a known continuous model can assist in avoiding these sources of error during design of both the model and its implementation, rather than relying on a potentially more arduous process of trial and error testing. Selection of an appropriate discretization can also maximize the model's ability to discriminate position by maintaining a sufficiently sharp approximation to the actual functions. Serious undersampling can produce much smoother, aliased approximations that discard potentially useful distinctions between positions.

The model assumed in order to apply Kalman filters does incorporate continuous state variables with observations and can be modified to accommodate changes in elapsed time between observations. The linear model relating state variables at one time point to state variables at a previous time point could be used to define a convolution model over score position. However, the distributions must be assumed normal to permit use of a Kalman filter. Recall that this assumption simplifies the calculation and enables it to be done in closed form over a few parameters only. Unfortunately, the actual distributions used in this work do not satisfy this requirement. The distance distribution for musical performances is not accurately approximated by normal curves over short elapsed times. Lognormal distributions are used instead. Convolution of these distributions with arbitrary distributions over previous score position is not solvable in a closed form. In addition, observations such as pitch are clearly not well-modeled as linear transformations of score position. Furthermore, the sharp edges in observation distributions have been shown to be important to distinguishing actual position. These sharp changes in density provide the information to distinguish score positions. Consequently, although the numerical techniques actually used to implement the score-following model are computationally more expensive, they are worthwhile if the approximations necessary to support a closed form calculation would produce less accurate estimates.

Now as previously shown, multiplication of several density functions for distinct types of observations can yield a sharper observation distribution. If sufficient observation types were identified such that their joint distribution could be guaranteed always to yield something sufficiently well-approximated by a normal curve and to make the skewness of the distance distribution irrelevant, then the Kalman filter might become applicable for stochastic score following. The observation distributions could be multiplied numerically and subsequently approximated by a normal curve, implicitly defining a linear relationship between the position and the observation (though the exact relationship might change each time the filter is applied). However, based on the experiences of this work, successful application of Kalman filters probably would require impeccable models of the observation distributions and a large number of such near-perfect distributions. Otherwise, the joint observation distribution sometimes will not be unimodal with small variance. As already mentioned, it is important to represent the exact shape of this distribution with high accuracy.

The requirements for score following benefit from the use of a statistical model that differs from the previously described models. Similarly, the position estimation task differs from the previously mentioned applications in certain important ways. First, a comprehensive knowledge of the variables relevant to the desired solution and analytical models relating those variables are not available. This situation contrasts with many problems commonly addressed in operations research and signal processing where analytical models exist or all relevant variables are assumed to be known. Second, score following is a real-time application of statistical estimation during a process, not a static analysis of a process or a post-process analysis. The real-time requirements severely limit the amount of computation done to produce each estimate, simultaneously restricting the kind of signal analysis done to produce observations.

The real-time requirements for score following are far more stringent than for many of the other tasks mentioned. For instance, in the case of autonomous navigation by office robots as considered by Koenig (1996), position estimation for action planning does not demand reports at a high frequency. Also, all is not lost if the robot pauses momentarily to consider the alternatives or take additional sensor readings. Position estimation for real-time control of the robot's motion might have timing requirements more similar to those of score following. The segmentation of Shuttle data as addressed by Chrisman (1996) was not done in real time, though the utility of a real-time solution might be argued. Automated speech recognition for applications such as dictation are also less demanding. It is not necessary for the recognition system to correctly identify every word in a sentence the instant it is spoken or even immediately following completion of the sentence. The system can remain useful even if the recognition delay is perceivable. A tutoring system that listens to a student read aloud and corrects any errors (Mostow and Aist 1997; Mostow *et al*. 1994) can benefit from low latency speech recognition, enabling the system to correct students before they proceed too far beyond the error. However, timing requirements for even this task are likely to be much more lax than for musical performance.

Score position estimation also differs from many of the previously mentioned applications due to the high accuracy and precision necessary for estimates to be useful at all. For instance, suppose that a score-following system's position estimates are acceptably close to the actual position of the performer 99% of the time, introducing a recognizable accompaniment glitch 1% of the time. A piece containing 100 notes would contain no accompaniment errors with probability $0.99^{100} = 0.366$. No competent performer would enjoy rehearsing with a system that makes a noticeable mistake in 6 out of every 10 performances (even if it is able to recover) and would never dream of going on stage with such a system. In contrast, if a dictation system misrecognizes or omits 1 out of every 100 words, forcing the user to repeat a sentence or otherwise correct the system, it may still be useful with respect to document preparation. Similarly, if an office robot makes a wrong turn 1% of the time, delaying its arrival at the destination but not causing it to become completely lost, the robot may still be useful. Now the high accuracy and precision requirements for score position estimation are mildly tempered by the availability of a highly informative score and, in the case of accompanying live performances, by the singer's ability to respond to the accompaniment. The score contains information about expected duration and sequencing of notes, pitch and diction. However, highly accurate score following requires integration of all this information, which in turn necessitates extensive signal processing, statistical modeling and estimation. Furthermore, performers are not unlikely to deviate from the literal score.

One additional computational technique for solving stochastic estimation problems is worth mentioning, especially since it can be used for on-line estimation problems. Dynamic programming can be used to find an optimal n-step path (minimum or maximum cost path) between any pair of points in a set of points given a nonnegative cost for making a direct transition between any two points. For instance, if the points represent cities and costs are provided for the length of a section of road directly connecting each pair of cities (with no cities intervening), dynamic programming can be used to determine the shortest (or longest) path connecting any pair of cities using exactly *n* sections of road. Dynamic programming is often efficient because it decomposes a computation into a recursive formulation whereby solutions for the *n*-step version can be computed using solutions for a version of the same problem with only *n*—1 steps. This property is often ideal for on-line applications where several successive versions of the same problem must be solved, each version specifying an increasing value of *n*. Dynamic programming is an important tool in operations research.

Dynamic programming is also applicable to problems where the minimization or maximization is over a product of costs rather than a sum. For instance, a Markov model of a state-transition process assumes that the probability of each state is conditioned only on the preceding state. Consequently, the probability for any sequence of states can be decomposed as follows:

$$P[S_t = s_i, S_{t-1} = s_j, S_{t-2} = s_k, \ldots, S_1 = s_m] = P[S_t = s_i \,|\, S_{t-1} = s_j]\, P[S_{t-1} = s_j \,|\, S_{t-2} = s_k] \cdots P[S_1 = s_m]$$

Viewing the states as points and the transition probabilities as costs, dynamic programming can be used to determine the most likely sequence of states given *n* transitions of the process. Note that for a Markov model, the last state of the most probable state sequence after *n* transitions may not be identical to the most likely state after *n* transitions, depending upon the transition probabilities. Identifying the latter state requires computing the distribution over the states after *n* transitions. This distribution is equivalent to the marginal distribution describing the state after *n* transitions as calculated over all possible state sequences containing *n* transitions.

Similarly, hidden Markov models assume that the probability of each state is conditioned only on the preceding state and that the observations are conditioned only on the current state. The probability for any sequence of states and observations can be decomposed as follows:

$$P[V_t = v_i, S_t = s_i, V_{t-1} = v_j, S_{t-1} = s_j, \ldots, V_1 = v_1, S_1 = s_1] =$$

$$P[V_t = v_i \,|\, S_t = s_i]\, P[S_t = s_i \,|\, S_{t-1} = s_j] \cdots P[V_1 = v_1 \,|\, S_1 = s_1]\, P[S_1 = s_1]$$

The sequence of conditional probabilities on the right can be grouped into pairs consisting of one observation probability and one transition probability, both probabilities based upon a common state. Viewing state-observation pairs as points and the product of the probabilities in each pair as costs, dynamic programming can be used to determine the most likely sequence of states and observations given *n* observations of the process. If a specific sequence of *n* observations is available, the considered sequences of states and observations can be restricted to those containing the actual observation sequence, and dynamic programming can still be used to determine the most likely sequence. This operation is equivalent to finding the state sequence that maximizes the following conditional probability for a specific sequence of observations:

$$P[S_t = s_i,\ S_{t-1} = s_j, \ldots,\ S_1 = s_1 \,|\, V_t = v_{i,} V_{t-1} = v_j, V_1 = v_1]$$

$$= \frac{P[V_t = v_{i,} S_t = s_i,\ V_{t-1} = v_j,\ S_{t-1} = s_j, \ldots,\ V_1 = v_1,\ S_1 = s_1]}{P[V_t = v_1,\ V_{t-1} = v_j,\ \ldots, V_1 = v1]}$$

$$= \frac{P[V_t = v_i \,|\, S_t = s_i]\ P[S_t = s_i \,|\, S_{t-1} = s_j] \cdots P[V_1 = v_1 \,|\, S_1 = s_1]\ P[S_1 = s_1]}{P[V_t = v_1,\ V_{t-1} = v_j,\ \ldots, V_1 = v1]}$$

Thus the states in the solution define the most likely state sequence given the actual observations. Applying dynamic programming to determine this state sequence is often referred to as the *Viterbi algorithm*.

Prior to the use of hidden Markov models for automated speech recognition, a technique known as *dynamic time-warping* was popular for recognition systems. Dynamic time-warping is an example of dynamic programming applied to optimize the pairing of two distinct sequences—one newly observed instance of speech and one prepared example, a *template*. A template could be a unique instance of speech or possibly an average across several instances. As just discussed, dynamic programming can be applied to generate a probabilistic measure of optimality under certain assumptions. However, the original use of dynamic time-warping in speech recognition was based on minimizing a distance measure between the observed sequence and one or more templates. These distance measures generally constituted some measure of spectral distortion, comparing spectral representations from the new speech example and the templates. In the case of a template formed by averaging multiple instances of speech, such distance measures only indicate how far the observations are from the expected or average values. They do not quantify the variation commonly observed and do not provide a measure of the likelihood of observing such deviations. Several ornamented approaches to dynamic time-warping have been used. Typically, path constraints are included, such as forcing the sequence of template points in the alignment to be monotonically nondecreasing or limiting the number of points that are skipped over by each match in the alignment. Often these extensions attempt to encode heuristic assumptions and criteria for ensuring a reasonable alignment.

The score-following system does not model the distribution over paths through the score, only the distribution over the current position of the performer. Unlike automated speech recognition where the desired final answer equates to a legal path through the states, the objective for score following was defined as maximizing the likelihood of estimating the correct score position of the singer. Furthermore, since score position is represented as a continuous variable, the precise interpretation of this goal is to maximize the likelihood of estimating a position within 50 ms of the singer's actual position. This approach is used since the probability of any individual score position must be 0. One might consider path optimization as an alternative approach. In this case, the score-following system could attempt to estimate the most likely path of the singer through the score, where a path is defined as a sequence consisting of one score position estimate for each reported observation. However, since there are an infinite number of possible score positions for each discrete observation, there are also an infinite number of possible paths. Consequently, the probability of any individual path must be zero.

One could address this problem by attempting to estimate a sequence of most likely position regions for each reported observation, rather than a single point. These regions in combination would represent the most likely set of paths for the singer. A set of possible paths defined by a sequence of such regions will be referred to as a *trajectory*. The span of the regions could be limited to a consistent size, say 100 ms each. If an optimal path estimate were constructed by selecting the center of each individual region, then the optimization process could be interpreted as identifying the path most likely to be always within 50 ms of the singer's actual position. Note that this operation is defined for a path of length 3 by the following expression:

$$\underset{i,j,k}{\operatorname{argmax}} \int_{x=i-50}^{i+50} \int_{y=j-50}^{j+50} \int_{z=k-50}^{k+50} f_{IJK}(x,y,z)\, \partial z\, \partial y\, \partial x$$

where i, j and k represent score positions at different points in time. In general, such an estimate may not be identical to the sequence of individual score position estimates as generated by the score-following system already presented. This situation can occur even when each position is assumed independent of all positions earlier than the immediately preceding position, allowing the expression for the optimal path of length 3 to be written as follows:

$$\underset{i,j,k}{\operatorname{argmax}} \int_{x=i-50}^{i+50} \int_{y=j-50}^{j+50} f_{I|J}(x|\,y) \int_{z=k-50}^{k+50} f_{JK}(y,z)\, \partial z\, \partial y\, \partial x$$

Recall that the same situation occurs when estimating paths using Markov models.

Dynamic programming can be used in the continuous case of optimal trajectory estimation as well as for discrete Markov models. Consider that under the previously stated independence assumption, the following equality holds:

$$\max_{i,j,k} P\,[(i-50 \le I \le i+50) \wedge (j-50 \le J \le j+50) \wedge (k-50 \le K \le k+50)] =$$

$$\max_{i,j} \left\{ P\,[i-50 \le I \le i+50 | j-50 \le J \le j+50] \max_{k} P\,[(j-50 \le J \le j+50) \wedge (k-50 \le K \le k+50)] \right\}$$

Note that the maximum over the variable k can be viewed as a function that, for each possible value of j, defines the maximum value of the contained expression over all k. Now a *conditional cumulative distribution function*, or *conditional cdf*, is defined as follows:

$$F_{I|J}(i|j) = \int f_{I|J}(i|j)\,\partial i$$

Note that this function specifies the probability that *I* will assume a value at or below *i* for any given value of *j* and is obtained directly by integrating over the conditional density. Using conditional cdf's, our original probability statement can be rewritten as follows:

$$\max_{i,j,k} P\,[(i-50 \le I \le i+50) \wedge (j-50 \le J \le j+50) \wedge (k-50 \le K \le k+50)]$$
$$= \max_i \max_j \left(\int_{x=j-50}^{j+50} [F_{I|J}(i+50|x) - F_{I|J}(i-50|x)]\,\partial x\right) \times$$
$$\left[\max_k \left(\int_{y=k-50}^{k+50} [F_{J|K}(j+50|y) - F_{J|K}(j-50|y)]\,\partial y\right) \times P[k-50 \le K \le k+50]\right]$$

When considering this equation in combination with the following equality:

$$\max_{j,k} P\,[(j-50 \le J \le j+50) \wedge (k-50 \le K \le k+50)] =$$
$$\max_j \left[\max_k \left(\int_{y=k-50}^{k+50} [F_{J|K}(j+50|y) - F_{J|K}(j-50|y)]\,\partial y\right) \times P[k-50 \le K \le k+50]\right]$$

a common computation in both expressions becomes explicit. This presence of a shared computation required for multiple, successive solutions enables the use of dynamic programming when incrementally extending the optimal trajectory. Note that the common computation involves calculating a conditional cdf over score position whenever a new observation is received. This calculation produces a full conditional distribution over *i* and *j*. If this function is permitted to be an arbitrary distribution function, then numerical techniques must be applied and the conditional cdf's must also be sampled, just like all the other functions in the presented score-following model. However, directly computing $F_{I|D,V,J}$ for all values of *J* and *I* would be considerably more expensive than computing $f_{I|D,V}$ for fixed values of *D* and *V*, as done in the score-following model presented. Determining whether or not accurate, real-time computation of an optimal trajectory is feasible is beyond the scope of this work, but may be worthy of further investigation.

Finally, it is worth mentioning one other statistical modeling technique that has recently become popular. *Bayesian networks* may be defined as a graphical interpretation of conditional probability. The graphs are directed and acyclic, with each node of the graph representing usually one but sometimes several possible events. Each node is associated with a conditional probability distribution over the events, with the events contained in all parent nodes serving as conditioning values. Bayesian networks are often used for incremental, uncertain symbolic reasoning applications and generally involve discrete variables in conditional distributions. As new information is made available, the conditional probability distributions are updated by propagating Bayes' rule over the graph. The appeal of the graphical representation is the ability to directly apply graph algorithms for computation. However, efficiency of the algorithms often requires that the number of arcs in the graph be restricted, implying many assumptions of independence. Consequently, this approach is better designed for executing simultaneous updates of large numbers of estimated variables.

In summary, there are several key differences between the methods applied in this work and the other statistical techniques, models, and algorithms described in this section. With respect to dynamic time warping, probability measures can be more informative than distance metrics. The former provides an estimate of expected variation in addition to simply the exact difference from the expected or average value. Methods for empirically estimating probabilities also exist, often allowing them to accurately approximate observed behavior in the real-world. With respect to discrete models, consideration of continuous time and continuous state variables can be important. In addition, it is often beneficial to consider the continuous distributions before developing a discrete model or implementation. This sequence of model development may enable one to maximize the discriminatory ability of the available variables and to minimize the potential approximation errors. Also, it can be beneficial to apply a continuous representation of time rather than a discrete representation, particularly when observations are not guaranteed to be available at regular intervals. Finally, numerical techniques for approximating continuous distributions can permit integration of arbitrary continuous and discrete distributions without precluding the use of more mathematically well-behaved, continuous distributions. Thus, the latter distributions can be used to simplify modeling and estimation when possible without forcing universal application of closed-form distributions and possibly introducing an unacceptable source of approximation error.

Although each of the described techniques can be appropriate for different applications exhibiting different requirements, the open framework applied in this project can be helpful. It is particularly appropriate when the available data is limited, a number of different variables must be integrated to produce estimates, at least some of the variables are continuous, accuracy of the estimates is a concern, and incremental development of the model and implementation is expected. The act of developing and estimating the model requires some proficiency with probability and statistics, but forces direct consideration of a model's assumptions relative to the application's objectives. The presented stochastic score-following model is not ideal for the performance tracking task; it can be computationally intensive and certain of the assumptions can be violated. However, it supports consideration of estimation accuracy relative to the targeted application and facilitates incremental development to improve the tracking. This approach may be appropriate for solving statistical estimation problems in other software applications, providing useful guidelines for incorporating multiple variables, assumptions, and solution methods that are most appropriate for each specific application.

8.3 Conclusions and Contributions

The work undertaken for this project has made several contributions to the understanding of both automated score following for vocal performances and statistical modeling for development of robust software. First, the most direct product of this effort is a robust and general system for tracking vocal performances. Although the system does not appear to achieve listening ability equivalent to human competency (based on available numbers), the tracking accuracy is often sufficient to enable reasonable performance of the accompaniment. In addition, this accuracy is achieved across many performances by skilled singers, these performances including different singers, pieces, styles and genres. The only required alterations specific to each performance (besides a symbolic score of the sung piece) involve configuring parameters for signal processing and tempo estimation.

Specification of the stochastic score-following model has necessitated collection, quantification, and synthesis of many properties of vocal performances. In particular, it was necessary to enumerate specific properties that complicate the tracking and analysis of vocal performances. These efforts required integration of prior knowledge about pitch detection, onset detection, and spectral envelope. It included work done in a variety of disciplines and subareas, including acoustics, computer music, vocal performance, music perception, and automated speech processing and recognition. It was also necessary to define distributions that describe characteristics of vocal performances and to identify metrics for assessing tracking accuracy and synchronization.

Implementation of the stochastic score-following model has applied general numerical techniques for computing distributions. This method is in contrast to both closed-form calculation techniques that limit the usable distributions and numerical approaches that do not directly support domain-specific modeling. It is based upon first defining component functions in the most appropriate form and then applying uniform sampling at the necessary resolution (just as in signal processing). This approach may be preferable to using a nonuniform or arbitrary discretization. Specifically, it can simplify analysis and modeling without sacrificing accuracy of the estimated distributions, especially compared to state-based models that limit the possible distributions. This point is particularly relevant to value estimation problems where the most accurate possible models are continuous and the errors introduced by discretization must be examined. Although nonuniform discretization may offer some improvement with respect to computational efficiency, the difference in computing time is probably not a major concern. In order to achieve equivalent or even acceptable precision and calculation accuracy, techniques based on nonuniform discretization probably require a number of samples of the same order of magnitude as the number of samples needed when applying uniform sampling. Also, since doubling of the sample rate yields a noticeable increase in both calculation accuracy and precision of the estimated value, any advantages to using nonuniform sampling for real-time (bounded time) computation will diminish over time so long as available computational power doubles at a fairly rapid pace.

This project has successfully applied probabilistic modeling to a problem for which it is not possible to specify a deterministic function that maps available inputs to the desired value (in this case the score position of a vocalist). The modeling proceeded via a top-down approach whereby the desired distribution is approximated by operations over component distributions that are easier to define and estimate. Use of appropriate notation and analysis techniques enabled a detailed consideration of all the assumptions and approximations necessary to produce a final decomposition that is both tractable to estimate and feasible to compute in real time. This approach facilitated statistically sound integration of multiple observations and explicit, continuous representations of time and rate. This integration included developing a continuous model of a performer's motion through the score assuming a positive-valued rate, and improving this model so that it produces consistent distributions even when the elapsed time between model updates varies.

As part of defining distributions for multiple observation types and evaluating the improvements in tracking accuracy, an informal statistical explanation was developed for why use of multiple observation types (*i.e.*, adding variables to the model) can lead to enhanced prediction (small error variance) and therefore improved tracking accuracy. Such improvement can occur only if the individual observation types provide useful discrimination of score positions, and the distributions for the observation types are estimated accurately and include the relevant factors influencing the observations. Methods for assessing the degree of improvement also were presented. These methods include a closed-form calculation, the expected likelihood of confusion, that is based on the content of musical scores and properly estimated distributions for the defined observation types. This value can be used as a preliminary assessment of expected average improvement in tracking accuracy when extending the stochastic score-following model, possibly expediting incremental enhancement of the model by reducing reliance on live performance testing.

Beyond the presented stochastic model for robust score following, this work has outlined and applied a methodology for incremental development and enhancement of performance tracking and accompaniment control software. More generally, this approach can be viewed as a methodology for incremental development and enhancement of statistical models for estimating desired variables. The applied methodology is not reliant on comprehensive data for defining a complex conditional distribution and does not require exclusive use of automated, black-box model optimization techniques. It facilitates use of analytical techniques where applicable and preferable, as well as the incorporation of simplifying assumptions when data collection just is not feasible. In addition, the compositional nature of the modeling permits for retraction of simplifying assumptions when sufficient data does become available or better analytical models are identified. This general framework may subsume use of models described in the preceding section, assisting in their integration and extension. If the automated optimization of arbitrary models is feasible for approximating component distributions, then these techniques can be incorporated within the overall estimation model as an approximation to one or more component density functions.

The overall objective of the statistical modeling is to "design" the distributions to improve accuracy of the estimated values. This goal is achieved by selecting the right information for use in estimating the predicted values and by minimizing the possible error sources through careful choice of assumptions, good estimation of distributions, and proper numerical analysis and implementation. By "right information" is meant the useful conditioning variables for every conditional distribution that appears in the model equations. Furthermore, correct values for these variables must be available when producing a new estimate of the predicted variable. This modeling process can be used to reduce the variance of the predictions to an acceptable level through incremental improvement and development. As this variance is reduced, the distributions will become closer to a true, deterministic function mapping the available information to an estimate of the desired value.

Future work on automated accompaniment systems should continue to investigate stochastic score following. In particular, identifying ways to sharpen observation distributions would be valuable, both in terms of the important, individual observation types considered in this project and the joint distribution over all available observation types. Considering enhanced sensors for the observation types already investigated, as well as for novel observation types, would be worthwhile. Examining more precise analytical models for both observation and distance distributions is another alternative. Identifying additional conditioning variables upon which to base the distributions might improve tracking accuracy as well. Stochastic score-following models could be applied to track performances by other instruments that are difficult to follow using acoustic information, especially stringed instruments such as violins.

Another interesting possibility is the automatic refinement of both the distributions and the score during rehearsal with the same performer. In automated speech recognition, training of stochastic models by individual users prior to recognition is a well-known and commonly applied technique. Work by Vercoe and Puckette (1985) applied simple statistical techniques to identify recurring, unmarked tempo changes made by the same musician performing the same piece. The vocal system by Inoue and colleagues (1994) required singers to perform each of the five Japanese vowels prior to tracking actual performances. Investigating similar techniques applied to the diverse components of the stochastic tracking method could lead to enhanced tracking through rehearsal. Modifying distributions to improve tracking would of course depend upon the accuracy of the original estimates. However, re-estimating positions in a post-performance mode would allow the system to incorporate all observations when determining each position estimate, including those that became available only after an estimate had been produced during real-time performance. Simultaneous estimation of all point estimates (a trajectory for the full performance) might also be feasible. Such post-processing techniques might provide position estimates that are more reliable than those available in real time. Comparing differences between the two estimates might indicate regions of the score that are especially problematic for automated tracking.

In addition, the retained connection between the estimated distributions and an informative symbolic score could be used to support an advanced user interface. For instance, if post-performance analysis could be used to identify possible estimation errors during the performance, the accompaniment system could request feedback from the user and focus discussion on the specific instances in question. This interaction might include assisting the user to modify the score, indicating tempo changes, differences in diction, and any ornamentation applied to the performance.

Finally, the accompaniment control mechanism applied in this project is very simple, and a rigorous, comprehensive investigation of how to accompany a singer has not been undertaken. It would be useful to develop a more precise, detailed description of how to control accompaniment to achieve a desired level of musicality. Important points to focus on would include accurately identifying places in the score where the accompanist should lead rather than follow. Knowing when to resume tempo after a *ritard* and what tempo to select is critical to providing an aesthetic performance. Extending the work of Mecca (1993), discovering how human accompanists perform, could be worthwhile. Precise quantification of how to vary the tempo and respond to the live performer is probably necessary.

Developing software for real-time, interactive musical performance is a demanding task offering many interesting and challenging problems. For this project, an attempt was made to focus on one particular component of one instance of a musical performance task—the robust, real-time tracking of vocal performances. In reality, it was necessary to widen this focus. Both the design and the evaluation of a score-following system have required recognition and examination of several problems and related tasks. Full and proper resolution of all the issues important to automated performance will likely require many volumes of many works, each as extensive as the present document. In short, many interesting opportunities remain.

Bibliography

Aitchison, J. and Brown, J.A.C. 1957. *The Lognormal Distribution, with Special Reference to Its Uses in Economics*. Cambridge: University Press.

Anderson, W.J. 1991. *Continuous-time Markov Chains: An Applications-oriented Approach*. New York: Springer-Verlag New York, Inc.

Atal, B.S. and Hanauer, S.L. 1971. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. In *Journal of the Acoustical Society of America* 50(2):637-55.

Austin, S., Barry, C., Chow, Y.-L., Derr, A., Kimball, O., Kubala, F., Makhoul, J., Placeway, P., Russell, W., Schwartz, R. and Yu, G. 1989. Improved HMM Models for High Performance Speech Recognition. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 249-255. San Mateo, CA: Morgan Kaufmann, Publishers, Inc.

Bahl, L.R., Jelinek, F. and Mercer, R. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5(2):179-90.

Bahl, L.R., Brown, P.F., deSouza P.V. and Mercer, R.L. 1987. Speech Recognition with Continuous-parameter Hidden Markov Models. In *Computer Speech and Language* 2(3-4):219-34.

Baird, B., Blevins, D. and Zahler, N. 1989. The Artificially Intelligent Computer Performer on the McIntosh II and a Pattern Matching Algorithm for Real-time Interactive Performance. In *Proceedings of the 1989 International Computer Music Conference*, 13-16. San Francisco: International Computer Music Association (ICMA).

Baird, B., Blevins, D. and Zahler, N. 1993. Artificial Intelligence and Music: Implementing an Interactive Computer Performer. In *Computer Music Journal* 17(2):73-9.

Baker, J.K. 1975. The DRAGON System—An Overview. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23(1):24-9.

Bartholomew, W.T. 1934. A Physical Definition of "Good Voice Quality" in the Male Voice. In *Journal of the Acoustical Society of America* 6(1):25-33.

Baum, L.E. and Petrie, T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. In *Annals of Mathematical Statistics* 37:1554-63.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. In *Annals of Mathematical Statistics* 41(1):164-71.

Beaulieu, N.C., Abu-Dayyal, A.A. and McLane, P.J. 1995. Estimating the Distribution of a Sum of Independent Lognormal Variables. In *IEEE Transactions on Communications* 43(12):2869-73.

Benade, A.H. 1976. *Fundamentals of Musical Acoustics*. New York: Oxford University Press.

Bilmes, J. 1992. A Model for Musical Rhythm. In *Proceedings of the 1992 International Computer Music Conference*, 207-10. San Francisco: International Computer Music Association (ICMA).

Bloch, J. and Dannenberg, R.B. 1985. Real-time Computer Accompaniment of Keyboard Performances. In *Proceedings of the 1985 International Computer Music Conference*, 279-90. San Francisco: International Computer Music Association (ICMA).

Bracewell, R.N. 1986. *The Fourier Transform and Its Applications*. 2nd ed. New York: McGraw-Hill.

Brigham, E.O. 1974. *The Fast Fourier Transform*. Englewood Cliffs, NJ: Prentice-Hall.

Broida, T.J. and Chellappa, R. 1986. Estimation of Object Motion Parameters from Noisy Images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(1):90-9.

Brown, J.C. 1991. Calculation of a Constant Q Spectral Transform. In *Journal of the Acoustical Society of America* 89(1):425-34.

Brown, J.C. 1992. Musical Fundamental Frequency Tracking Using a Pattern Recognition Approach. In *Journal of the Acoustical Society of America* 92(3):1394-1402.

Casajus-Quiros, F.J. and Fernandez-Cid, P. 1994. Real-time, Loose-harmonic Matching Fundamental Frequency Estimation for Musical Signals. In *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing*, 221-4. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Charpentier, F.J. 1986. Pitch Detection Using the Short-term Phase Spectrum. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 113-6. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Chrisman, L.D. 1996. Approximation of Graphical Probabilistic Models by Iterative Dynamic Discretization and its Application to Time-series Segmentation. Doctoral dissertation, Department of Computer Science, Carnegie Mellon University, Pittsburgh.

Clarke, E. 1989. The Perception of Expressive Timing in Music. *Psychological Research* 51:2-9.

Coffin, B., *et al.* 1982. *Phonetic Readings of Songs and Arias*. 2nd ed. Metuchen, New Jersey: Scarecrow Press.

Colorni, E. 1996. *Singers' Italian: A Manual of Diction and Phonetics*. New York: Schirmer Books.

Cooper, D. and Ng, K.C. 1996. A Monophonic Pitch-tracking Algorithm Based on Waveform Periodicity Determinations Using Landmark Points. In *Computer Music Journal* 20(3):70-8.

Cowan, C.F.N. and Grant, P.M. 1985. *Adaptive Filters*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Cox, R.G. 1970. *The Singer's Manual of German and French Diction*. New York: G. Schirmer, Inc.

Crowley, J.L. 1995. Mathematical Foundations of Navigation and Perception for an Autonomous Mobile Robot. In *Lecture Notes in Artificial Intelligence: Reasoning with Uncertainty in Robotics*, vol. 1093, 9-51. Berlin: Springer-Verlag.

Dannenberg, R.B. 1984. An On-line Algorithm for Real-time Accompaniment. In *Proceedings of the 1984 International Computer Music Conference*, 193-8. San Francisco: International Computer Music Association (ICMA).

Dannenberg, R.B. 1986. The CMU MIDI Toolkit. In *Proceedings of the 1986 International Computer Music Conference*, 53-6. San Francisco: International Computer Music Association (ICMA).

Dannenberg, R.B. 1993a. *The CMU MIDI Toolkit*. Pittsburgh: Carnegie Mellon University.

Dannenberg, R.B. 1993b. Software Design for Interactive Multimedia Performance. In *Interface—Journal of New Music Research* 22(3):213-28.

Dannenberg, R.B. and Bookstein, K. 1991. Practical Aspects of a MIDI Conducting Program. In *Proceedings of the 1991 International Computer Music Conference*, 537-40. San Francisco: International Computer Music Association (ICMA).

Dannenberg, R. and Mukaino, H. 1988. New Techniques for Enhanced Quality of Computer Accompaniment. In *Proceedings of the 1988 International Computer Music Conference*, 243-9.

Dautrich, B., Rabiner, L. and Martin, T. 1983. On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 31(4):793-807.

Davis, P. and Rabinowitz, P. 1967. *Numerical Integration*. Waltham, MA: Blaisdell Publishing Company.

Davis, S.B. and Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28(4):357-66.

Deng, L., Lenning, M. and Mermelstein, P. 1989. Use of Vowel Duration Information in a Large Vocabulary Word Recognizer. In *Journal of the Acoustical Society of America* 86(2):540-8.

Derrouault, A.-M. 1987. Context-dependent Phonetic Markov Models for Large Vocabulary Speech Recognition. In *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing*, 360-3. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Desain, P. and Honing, H. 1992. Tempo Curves Considered Harmful. In *Music, Mind, and Machine: Studies in Computer Music, Music Cognition, and Artificial Intelligence*, ed. Desain, P. and Honing, H. 25-40. Amsterdam: Thesis Publishers.

Doval, B. and Rodet, X. 1993. Fundamental Frequency Estimation and Tracking Using Maximum Likelihood Harmonic Matching and HMMs. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 221-4. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Fernandez-Cid, P. and Casajus-Quiros, F.J. 1994. DSP Based Reliable Pitch-to-MIDI Converter by Harmonic Matching. In *Proceedings of the 1994 International Computer Music Conference*, 307-10. San Francisco: International Computer Music Association (ICMA).

Foster, S., Schloss, W.A. and Rockmore, A.J. 1982. Toward an Intelligent Editor of Digital Audio: Signal Processing Methods. In *Computer Music Journal* 6(1):42-51.

French, N.R. and Steinberg, J.C. 1947. Factors Governing the Intelligibitliy of Speech Sounds. In *Journal of the Acoustical Society of America* 19:90-119.

Fox, D., Burgard, W., Thrun, S. and Cremers, A. 1998. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. Menlo Park, CA: AAAI Press.

Furui, S. 1986. Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34(1):52-9.

Gray, Jr., A.H. and Markel, J.D. 1976. Distance Measures for Speech Processing. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24(5):380-91.

Gray, R.M. 1984. Vector Quantization. In *IEEE Acoustics, Speech, and Signal Processing Magazine* 1(2):4-29.

Grubb, L. and Dannenberg, R.B. 1994a. Automated Accompaniment of Musical Ensembles. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 94-9. Cambridge, MA: MIT Press.

Grubb, L. and Dannenberg, R.B. 1994b. Automating Ensemble Performance. In *Proceedings of the 1994 International Computer Music Conference*, 63-9. San Francisco: International Computer Music Association (ICMA).

Gupta, V.N., Lenning, M. and Mermelstein, P. 1987. Integration of Acoustic Information in a Large Vocabulary Word Recognizer. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 697-700. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Hermes, D.J. 1988. Measurement of Pitch by Subharmonic Summation. In *Journal of the Acoustical Society of America* 83(1):257-64.

Hirata, K., ed. 1997. *Working Notes from the IJCAI-97 Workshop on Issues in AI and Music—Evaluation and Assessment* (unpublished).

Hollien, H., Hollien, P. and deJong, G. 1997. Effects of Three Parameters on Speaking Fundamental Frequency. In *Journal of the Acoustical Society of America* 102(5):2984-92.

Horii, Y. 1989. Frequency Modulation Characteristics of Sustained /ɑ/ Sung in Vocal Vibrato. In *Journal of Speech and Hearing Research* 32(4):829-36.

Howard, R.A. 1971. *Dynamic Probabilistic Systems*. 2 vols. New York: John Wiley and Sons, Inc.

Huang, H.D. and Jack, M.A. 1989. Semi-continuous Hidden Markov Models for Speech Signals. In *Computer Speech and Language* 3(3):239-52.

Huang, H.D., Lee, K.F., Hon, H.W. and Hwang, M.Y. 1991. Improved Acoustic Modeling with the SPHINX Speech Recognition System. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 345-8. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Inoue, W., Hashimoto, S., and Ohteru, S. 1993. A Computer Music System for Human Singing. In *Proceedings of the 1993 International Computer Music Conference*, 150-3. San Francisco: International Computer Music Association (ICMA).

Inoue, W., Hashimoto, S., and Ohteru, S. 1994. Adaptive Karaoke System—Human Singing Accompaniment Based on Speech Recognition. In *Proceedings of the 1994 International Computer Music Conference*, 70-7. San Francisco: International Computer Music Association (ICMA).

Jelinek, F. 1976. Continuous Speech Recognition by Statistical Methods. In *Proceedings of the IEEE* 64(4):532-56.

Kalman, R.E. 1960. A New Approach to Linear Filtering and Prediction Problems. In *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering* 82 (March 1960):35-45.

Katayose, H., Kanamori, T., Kamei, K., Nagashima, Y., Sato, K., Inokuchi, S. and Simura, S. 1993. Virtual Performer. In *Proceedings of the 1993 International Computer Music Conference*, 138-145. San Francisco: International Computer Music Association (ICMA).

Kendall, M.G., Stuart, A. and Ord, K.J. 1983. *Kendall's Advanced Theory of Statistics*. 4th ed. 3 vols. New York: Oxford University Press.

Koenig, S. 1996. Goal-directed Acting with Incomplete Information. Doctoral dissertation, Department of Computer Science, Carnegie Mellon University, Pittsburgh.

Koenig, S. and Simmons, R. 1996. Unsupervised Learning of Probabilistic Models for Robot Navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2301-10. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Kosaka, A. and Kak, A. 1992. Fast Vision-guided Mobile Robot Navigation Using Model-based Reasoning and Prediction of Uncertainties. In *Image Understanding* 56(3):271-329.

Kuhn, W. 1990. A Real-time Pitch Recognition Algorithm for Music Applications. In *Computer Music Journal* 14(3):60-71.

Lee, C.-H., Rabiner, L.R., Pieraccini R. and Wilpon, J.G. 1989. Acoustic Modeling of Subword Units for Large Vocabulary Speaker Independent Speech Recognition. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 280-291. San Mateo, CA: Morgan Kaufmann, Publishers, Inc.

Lee, C.-H., Giachin, E., Rabiner, L.R., Pieraccini, R. and Rosenberg, A.E. 1992. Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition. In *Computer Speech and Language* 6(2):103-27.

Lee, K.-F. 1989. Hidden Markov Models: Past, Present, and Future. In *Proceedings of the European Conference on Speech Communication and Technology*, 148-55. Edinburgh, U.K.: CEP Consultants.

Lee, K.-F. 1990. Context-dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38(4):599-609.

Legters, Jr., G.R. and Young, T.Y. 1982. A Mathematical Model for Computer Image Tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(11):583-94.

Levinson, S.E. 1986. Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. In *Computer Speech and Language* 1(1):29-45.

Lifton, J. 1985. Some Technical and Aesthetic Considerations in Software for Live Interactive Performance. In *Proceedings of the 1985 International Computer Music Conference*, 303-6. San Francisco: International Computer Music Association (ICMA).

Linde, Y., Buzo, A. and Gray, R. M. 1980. An Algorithm for Vector Quantizer Design. In *IEEE Transactions on Communications* COM-28:84-95.

Maher, R.C. and Beauchamp, J.W. 1994. Fundamental Frequency Estimation of Musical Signals Using a Two-way Mismatch Procedure. In *Journal of the Acoustical Society of America* 95(4):2254-63.

Maher, R.C. and Beauchamp, J.W. 1990. An Investigation of Vocal Vibrato for Synthesis. In *Applied Acoustics* 30(2-3):219-45.

Markel, J.D. and Gray, Jr., A.H. 1976. *Linear Prediction of Speech*. Berlin: Springer-Verlag.

Marshall, M. 1953. *The Singer's Manual of English Diction*. New York: G. Schirmer, Inc.

Martin, J.J. 1967. *Bayesian Decision Problems and Markov Chains*. New York: John Wiley and Sons, Inc.

Maybeck, P.S. 1982. *Stochastic Models, Estimation and Control*. 3 vols. New York: Academic Press.

Mecca, M. 1993. Tempo Following Behavior in Musical Accompaniment. Master's thesis, Department of Philosophy, Carnegie Mellon University, Pittsburgh.

MIDI Manufacturers Association (MMA) and the Japan MIDI Standards Committee (JMSC). 1989. *MIDI 1.0 Detailed Specification*. Los Angeles: International MIDI Association.

Mitra, S.K. and Kaiser, J.F. 1993. *Handbook for Digital Signal Processing*. New York: John Wiley and Sons.

Monkowski, M.D., Picheny, M. and Srinivasan Rao, P. 1995. Context Dependent Phonetic Duration Models for Decoding Conversational Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 528-31. New York: IEEE.

Moriarty, J. 1975. *Diction: Italian, Latin, French, German; the sounds and 81 exercises for singing them*. 2nd ed. Boston: E.C. Schirmer Music Company.

Mostow, J. and Aist, G. 1997. The Sounds of Silence: Towards Automated Evaluation of Student Learning in a Reading Tutor that Listens. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 355-61. Menlo Park, CA: AAAI Press.

Mostow, J., Roth, S.F., Hauptmann, A.G. and Kane, M. 1994. A Prototype Reading Coach that Listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 785-92. Cambridge, MA: MIT Press.

Mulgrew, B. and Cowan, C.F.N. 1988. *Adaptive Filters and Equalisers*. Norwell, MA: Kluwer Academic Publishers.

Niihara, T., Imai, M. and Inokuchi, S. 1986. Transcription of Sung Song. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1277-80. New York: IEEE.

Noll, A. 1970. Cepstrum Pitch Determination. In *Journal of the Acoustical Society of America* 47(2):634-48.

Odom, W. 1981. *German for Singers: A Textbook of Diction and Phonetics*. New York: Schirmer Books.

Oppenheim, A.V. and Schafer, R.W. 1975. *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Oppenheim, A.V. and Weinstein, C. 1972. Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform. In *IEEE Proceedings* 60:957-76.

Paul, D.B. and Martin, E.A. 1988. Speaker Stress-resistant Continuous Speech Recognition. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing*, 283-6. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Pearl, J.D. 1982. Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-82)*, 133-6. Los Altos, CA: William Kaufmann.

Povel, D.-J. and Essens, P. 1985. Perception of Temporal Patterns. In *Music Perception* 2(4):411-40.

Prame, E. 1994. Measurements of the Vibrato Rate of Ten Singers. In *Journal of the Acoustical Society of America*, 90(4):1979-84.

Puckette, M. 1995. Score Following Using the Sung Voice. In *Proceedings of the 1995 International Computer Music Conference*, 175-8. San Francisco: International Computer Music Association (ICMA).

Rabiner, L. 1977. On the Use of Autocorrelation Analysis for Pitch Detection. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25(1):24-33.

Rabiner, L., Cheng, M., Rosenberg, A. and McGonegal, C. 1976. A Comparative Performance Study of Several Pitch Detection Algorithms. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24(5):399-418.

Rabiner, L. and Schafer, R. 1978. *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall.

Rabiner, L.R., Levinson, S.E. and Sondhi, M.M. 1983. On the Application of Vector Quantization and Hidden Markov Models to Speaker-independent Isolated Word Recognition. In *Bell System Technical Journal* 62(4):1075-1105.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE* 77(2):257-86.

Rossing, T.D., Sundberg, J. and Ternström, S. 1987. Acoustic Comparison of Soprano Solo and Choir Singing. In *Journal of the Acoustical Society of America* 82(3):830-6.

Russell, M.J. and Moore, R.K. 1985. Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 49-52. New York: The Institute of Electrical and Electronics Engineers (IEEE).

Schulze, H. 1989. Categorical Perception of Rhythmic Patterns. In *Psychological Research* 51:10-15.

Schwartz, S.C. and Yeh, S. 1982. On the Distribution Function and Moments of Power Sums with Log-normal Components. In *Bell System Technical Journal* 61(7):1441-62.

Simmons, R. and Koenig, S. 1995. Probabilistic Robot Navigation in Partially Observable Environments. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, 1080-7. San Mateo, CA: Morgan Kaufmann Publishers.

Simons, B., Welch, J.L. and Lynch, N. 1990. An Overview of Clock Synchronization. In *Fault-tolerant Distributed Computing*, ed. B. Simons and A. Spector, 84-96. Berlin: Springer-Verlag.

Smith, R.C. and Cheeseman, P. 1986. On the Estimation and Representation of Spatial Uncertainty. In *International Journal of Robotics Research* 5(4): 56-68.

Solbach, L. and Wöhrmann, R. 1996. Sound Onset Localization and Partial Tracking in Gaussian White Noise. In *Proceedings of the 1996 International Computer Music Conference*, 324-7. San Francisco: International Computer Music Association (ICMA).

Steiglitz, K., Winham, G. and Petzinger, J. 1975. Pitch Extraction by Trigonometric Curve Fitting. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23(3):321-3.

Sternberg, S., Knoll, R. and Zukofsky, P. 1982. Timing by Skilled Musicians: Perception, production and imitation of time ratios. In *The Psychology of Music*, ed. D. Deutsch, 181-239. New York: Academic Press.

Stevens, S.S. and Volkmann, J. 1940. The Relation of Pitch to Frequency: A Revised Scale. In *American Journal of Psychology* 53:329-53.

Sundberg, J. 1974. Articulatory Interpretation of the "Singing Formant". In *Journal of the Acoustical Society of America* 55(3):838-44.

Sundberg, J. 1975. Formant Technique in a Professional Female Singer. In *Acustica* 32:89-96.

Sundberg, J. 1987. *The Science of the Singing Voice*. DeKalb, Illinois: Northern Illinois University Press.

Sundberg, J. 1991. *The Science of Musical Sounds*. San Diego: Academic Press, Inc.

Tait, C. and Findlay, W. 1996. Wavelet Analysis for Onset Detection. In *Proceedings of the 1996 International Computer Music Conference*, 500-3. San Francisco: International Computer Music Association (ICMA).

Taylor, I. and Greenhaugh, M. 1993. An Object-oriented ARTMAP System for Classifying Pitch. In *Proceedings of the 1993 International Computer Music Conference*, 244-7. San Francisco: International Computer Music Association (ICMA).

Taylor, I. and Greenhaugh, M. 1994. Evaluation of Artificial-neural-network Types for Determination of Pitch. In *Proceedings of the 1994 International Computer Music Conference*, 114-20. San Francisco: International Computer Music Association (ICMA).

Thorin, O. 1977. On the Infinite Divisibility of the Lognormal Distribution. In *Scandinavian Actuarial Journal*, 1977:121-48.

Thrun, S., Burgard, W. and Fox, D. 1998. Concurrent Mapping and Loccalization for Mobile Robots. In *Machine Learning*, 31(1-3):29-53.

Uris, D. 1971. *To Sing in English: A Guide to Improved Diction*. New York: Boosey and Hawkes.

Vennard, W. 1967. *Singing—the Mechanism and the Technique*. 2nd ed. New York: Carl Fischer, Inc.

Vercoe, B. 1984. The Synthetic Performer in the Context of Live Performance. In *Proceedings of the 1984 International Computer Music Conference*, 199-200. San Francisco: International Computer Music Association (ICMA).

Vercoe, B. and Puckette, M. 1985. Synthetic Rehearsal: Training the Synthetic Performer. In *Proceedings of the 1985 International Computer Music Conference*, 275-8. San Francisco: International Computer Music Association (ICMA).

Zwicker, E., Flottorp, G. and Stevens, S.S. 1957. Critical Bandwidth in Loudness Summation. In *Journal of the Acoustical Society of America* 29:548-57.

Appendix

Musical Works Included in this Study

The following lists indicate all musical works in each of the three sets of performances collected during this study—the 20 performances used for estimating distributions, the 8 recorded performances used to evaluate the tracking system, and the 6 live performances used to evaluate the tracking system. No association between score and performance statistics are provided in an attempt to preserve anonymity among the performers. Note also that the number of works listed for each set of performances is smaller than the total number of performances since several singers performed the same piece.

The 20 recordings used for estimating distributions included performances of the following musical works:

"Widerstehe doch der Sünde" from cantata no. 54	J. S. Bach
"Simple gifts" from *Old American Songs*	A. Copland
"Why do they shut me out of Heaven?" from *Twelve Poems of Emily Dickinson*	A. Copland
"Who is Sylvia?" from *Let us Garlands Bring*	G. Finzi
"Come unto Him" from *Messiah*	G. F. Handel
"Dall'ondoso periglio" from *Giulio Caesare*	G. F. Handel
"Where'er you walk" from *Semele*	G. F. Handel
"Lullaby" from *The Consul*	G.-C. Menotti
"Un moto di gioia", K. 579	W. A. Mozart
"O mio babbino caro" from *Gianni Schicchi*	G. Puccini
"Vouchsafe, O Lord" from *Te Deum Laudamus*	H. Purcell
"Der Leiermann" from *Winterreisse*	F. Schubert
"Der Tod und das Mädchen"	F. Schubert
"Die Lotosblume"	R. Schumann
"A te l'estremo addio" from *Simon Boccanegra*	G. Verdi
"Swing low, sweet chariot"	spiritual

The eight recordings used for evaluating the tracking system included performances of the following musical works:

"Ecco mi in lieta veste" from *I Capuletti ed i Montecchi*	V. Bellini
"The boatmen's dance" from *Old American Songs*	A. Copland
"Cangio d'aspetto" from *Admeto*	G. F. Handel
"Si la voglio" from *Xerxes*	G. F. Handel
"Non più andrai" from *Le Nozze di Figaro*	W. A. Mozart
"Die Lotosblume"	R. Schumann
"Heimliche Aufforderung" from *4 Lieder für eine Singstimme*	R. Strauss

The six live performances included the following musical works:

"Glückwunsch" from *Fünf Lieder*	E. W. Korngold
"Bester Jüngling" from *Der Schauspieldirektor*	W. A. Mozart
"Non più andrai" from *Le Nozze di Figaro*	W. A. Mozart
"Sole e amore"	G. Puccini
"Lachen und Weinen"	F. Schubert