IMPROVING ACOUSTIC MODELS BY WATCHING TELEVISION

Michael J. Witbrock[2,3] and Alexander G. Hauptmann[1]

March 19th, 1998

CMU-CS-98-110

[1]School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA 15213-3890 USA

[2]Justresearch (Justsystem Pittsburgh Research Center),
4616 Henry St,
Pittsburgh PA 15213 USA

This work was first presented at the 1997 AAAI Spring Symposium, Palo Alto, CA., March 1997.

[3]The work described in this paper was done while the first author was an employee of Carnegie Mellon University

Abstract

Obtaining sufficient labelled training data is a persistent difficulty for speech recognition research. Although well transcribed data is expensive to produce, there is a constant stream of challenging speech data and poor transcription broadcast as closed-captioned television. We describe a reliable unsupervised method for identifying accurately transcribed sections of these broadcasts, and show how these segments can be used to train a recognition system. Starting from acoustic models trained on the Wall Street Journal database, a single iteration of our training method reduced the word error rate on an independent broadcast television news test set from 62.2 % to 59.5%.

This paper is based on work supported by the National Science Foundation, DARPA and NASA under NSF Cooperative agreement No. IRI-9411299. We thank Justsystem Corporation for supporting the preparation of the paper. The views and conclusions contained in this document are those of the authors and do not necessarily reflect the views of any of the sponsors.

Keywords: Digital Libraries, Speech Recognition, Alignment of Text and Speech, Speech Recogniser Training, Viterbi Search, Recognition Errors, Informedia.

Introduction

Current speech recognition research is characterized by its reliance on data. The statistical (Huang et al. 1994) and neural network (Kershaw, Robinson & Renals 1995) based recognizers that have become popular over the last decade depend on the automatic training of models with many thousands of parameters. These parameters can only be accurately estimated from large amounts of recorded speech; the slogan "there's no data like more data" is frequently heard in speech laboratories.

Fortunately, advances in storage technology and processing power have made the problem of managing huge quantities of data relatively simple, and the training process, while still a trial for the patience, at least tractable. Unfortunately, a great deal of effort must still be expended to *collect* the speech data itself, both in making careful recordings from suitable speakers, and in annotating the recordings with careful transcriptions. The work described in this paper takes a step towards reducing this cost by making use of large quantities of speech produced for other purposes.

The holy grail for speech training is a completely unsupervised system using an independent source of knowledge to detect and transcribe misrecognised or unknown words, thus allowing acoustic models to be reestimated. Lacking a complete solution, we have chosen to approach this goal by attempting unsupervised collection of training data.

Every day, vast quantities of speech are broadcast on television along with roughly corresponding closed-caption or teletext titles. As part of the Informedia project (Hauptmann & Witbrock 1996) at Carnegie Mellon, we have been capturing this speech, along with the broadcast captions, for use in a full-context digital video library retrieval system. These captions cannot, of course, be used directly. For broadcast news, our experiments have shown that approximately 16% of the words in closed captions are incorrectly transcribed when compared with careful transcripts of the same shows produced by the Journal Graphics, Inc. professional transcription service.

We use the Sphinx-II system, which is a large-vocabulary, speaker-independent, continuous speech recognizer created at Carnegie Mellon (CMU Speech 1997, Huang et al. 1994, Ravishankar 1996). Sphinx-II uses 10000 senonic semi-continuous hidden Markov models (HMMs) to model between-word context-dependent phones. Our language model was constructed from a corpus of news stories from the Wall Street Journal from 1989 to 1994 and the Associated Press news service stories from 1988 to 1990. Only trigrams that were encountered more than once were included in the model, along with all bigrams and the most frequent 20000 words in the corpus (Rudnicky 1995).

Our test data consisted of a thirty minute news show recorded independently from any of the training data. On this set, segmented into ninety "utterance" chunks, the recognition word error rate (substitutions+insertions+deletions) was 62.2%.

Analysis of the recognizer errors shows that even with a trigram language model derived from a correct transcript, there is a significant error rate (Placeway & Lafferty 1996). This leads to the conclusion that poor acoustic modeling is the major source of error for the broadcast television data. While Placeway and Lafferty used a particular closed-caption transcript as a hint to improve the recognition for the corresponding audio track, our purpose is to use closed-caption data to obtain a large correctly transcribed training corpus.

Previous work on automatic learning in speech recognition has focussed chiefly on unsupervised adaptation schemes. Cox and Bridle's connectionist RECNORM system (Cox & Bridle 1990), for example, improved recognition accuracy by simply training the recognition network to more confidently output its existing classification decisions. The HTK recogniser described in (Woodland et al. 1994) also used unsupervised speaker adaptation to improve accuracy.

Text Alignment of Speech Recognition and Closed Caption Data

The word error rate for the closed captions is high at 15.7%, but the baseline word error rate for the Sphinx II (Huang et al. 1994) recognizer applied to the test data is even worse: 62.2%. However, in using both of these sources to find the exact timings for word utterances on which Informedia depends, we have found that quite accurate text alignment between the speech recognition and the closed captions is possible. Since the errors made by the captioning service and those made by Sphinx are largely independent, we can be confident that extended sections over which the captions and the Sphinx transcript correspond have been correctly transcribed.The process of finding correspondences is rather straightforward: a dynamic programming alignment (Nye 1984) is performed between the two text strings, with a distance metric between words that is zero if they match exactly, one if they don't match at all, and which increases with the number of mismatched letters in the case of partial matches.

Once this method has found corresponding sections, it is a relatively simple matter to excerpt the corresponding speech signal and captioning text from their respective files, add them to the training set, and iterate. The effect of this process on recognition accuracy will be described later in the paper.

Because of the high error rates in the source material, only a small proportion of the words spoken can be identified as correct. The processing required to do this identification is not insignificant. The speech recogniser must be run on all the broadcast television audio. For the training experiments, a minimal acceptable span of three words was used, giving a yield of 4.5% of the spoken words (or, very approximately, 2.7 minutes of speech per hour of TV broadcast).

Training

The model for improving on acoustic models is quite simple, and is outlined in Figure 1

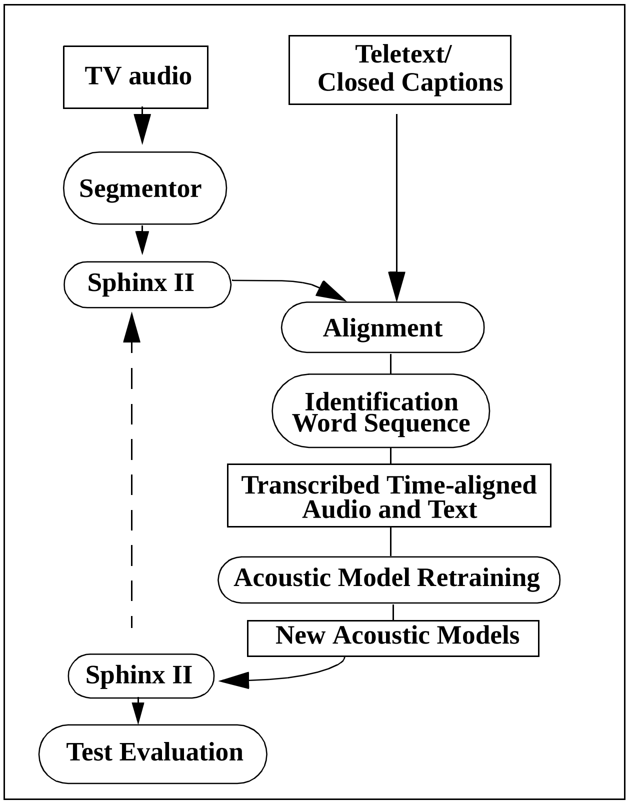

Figure 1: The process for retraining acoustic models based on television input. The Sphinx II speech recogniser is used along with the closed captions to identify a collection of segments where the transcript is accurate, and these segments are used to retrain acoustic models that can be used in subsequent iterations.

The closed-caption stream from the television is captured and time-stamped. At the same time, the audio track is captured and segmented into chunks on average thirty seconds long based on silence. Silence is defined as long, low energy periods in the acoustic signal. These chunks are then fed through the SPHINX-II speech recognition system running with a 20000 word vocabulary and a language model based on North American broadcast news from 1987 to 1994 (Rudnicky 1995). The recognition output for each chunk is aligned to the last few minutes of closed captions. If there are chunks of three or more contiguous words that match in the alignment, we assume a correct transcription. To avoid corrupting the transitions into the first word and out of the last word in the sequence, we remove the first and last words, since their acoustic boundaries might have been mis-characterized due to incorrectly recognized adhjacent words. Then we split out the audio sample corresponding to these words from the current chunk, and store it together with the transcribed words. At this point the transcription has been "verified" through two independent sources: The closed-caption text and the speech recognizer output. We can, therefore, be confident that the transcription is correct and can be used for adapting the current acoustic models. Examples of recognized phrases that we use for training are listed in Table 1.

The resulting data was then used to adapt the initial acoustic models. Initially, our acoustic models were derived from the Wall Street Journal training data (Huang et al. 1994), without distinction of gender. The adaptive training procedure (Sphinx III) was then used to modify the means and variances of the existing codebook entries according to the new training data. We did not retrain individual senone distributions, since we didn't have enough data to do so at that time. .

Table 1: Examples of well recognized segments identified by the alignment procedure. The segments used are ones for which the speech recogniser output and the closed captions agree for a span of more than three words.

the top royal according to a new
her estranged husband prince
to *SIL* share *SIL* even
SIL transplants from parents higher than from unrelated living donors *SIL*
white *SIL* house contends that
the republican strategy on *SIL*
questions today about his refusal to hand *SIL* over those
to turn over these notes
there is nothing extraordinary *SIL*
many *SIL* times in this

Results

The following results were derived from an initial run of the system. We expect to have more extensive data available in the next few months. 2987 training phrases were derived as described above. The phrases contained 18167 words (6.08 words per phrase). A total of 2948 distinct words were recognized from the maximal vocabulary of 20000 words in the speech recognition dictionary.

The baseline Word Error Rate (WER) is 62.2 % for the Sphinx-II system. Recognition accuracy improved to 59.5 % WER using the initial set of 2987 adaptation sentences that were automatically derived using the above described procedure.

Conclusions and Future Work

One possible criticism of the current scheme is that it identifies sections of speech on which the recognizer already works. It is to be hoped that there is sufficient variability in these sections to provide useful training, but it is possible that a plateau will be reached. One possibility for mitigating this effect is to accept single words in the captions that do not correspond to the SR output, providing that they are surrounded by correctly transcribed segments.

Despite the easy gains from a fairly small number of automatically selected phrases, several important questions remain at this point. One could argue that this technique will quickly reach an asymptote, since the speech recognition acoustic models are only adpating to what the speech recognizer already knows how to recognize. On the other hand, the recognizer bases it's recognition on both the acoustics as well as a static North American business news language model, so at times, poorly identified acoustics will be compensated for in the language model.

Another argument is that the initial fit of the acoustic models is so poor, that any minimal adaptation to the environment will result in an initial improvement. We hope to answer these concerns in the next few months of experimentation.

References

Cox, S.J., and Bridle, J.S. "Simultaneous speaker normalisation and utterance labelling using Bayesian/neural net techniques." 1990. In *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 1. pp 161-4.

Hauptmann, A. and Witbrock, M. 1996, Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval, In Maybury, M, ed, *"Intelligent Multimedia Information Retrieval"*, AAAI Press, Forthcoming.

Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., and Alleva, F., 1994, "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II." *ICASSP-94*, vol. I, pp. 549-552.

Kershaw, D.J. Robinson, A.J. and Renals, S.J., 1996, "The 1995 Abbot Hybrid Connectionist-HMM Large-Vocabulary Recognition System". In *Notes from the 1996 ARPA Speech Recognition Workshop*, *Arden House*, Harriman NY, Feb 1996.

Nye, H., 1984, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol AASP-32, No 2, pp 263-271

Rudnicky, A.I., 1996, "Language Modeling with Limited Domain Data," In *Proceeding of the 1995 ARPA Workshop on Spoken Language Technology.*

CMU Speech Group, 1997,
URL: http://www.speech.cs.cmu.edu/speech

Sphinx III Training, 1997,
http://www.cs.cmu.edu/~eht/s3_train/s3_train.html

Ravishankar, M. K., 1996, Efficient Algorithms for Speech Recognition, PhD diss. Carnegie Mellon University. Technical Report CMU-CS-96-143.

Placeway, P. and Lafferty, J., 1996, "Cheating with Imperfect Transcripts", In *Proceedings of ICSLP 1996*.

Woodland, P.C. Leggetter, C.J. Odell, J.J., Valtchev, V. Young, S.J., 1995, "The 1994 HTK large vocabulary speech recognition system", In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 1. pp 73-76.