

S1.9

Dynamic Adaptation of Hidden Markov Models for Robust Isolated-Word Speech Recognition¹

Edward A. Martin and Richard P. Lippmann and Douglas B. Paul

Lincoln Laboratory, M.I.T.
Lexington, MA 02173, USA

Abstract

This paper describes an HMM-based isolated-word recognition system that dynamically adapts word model parameters to new speakers and to stress-induced speech variations. During recognition all input tokens presented to the system can be used to augment the current word model parameters. New tokens can be weighted so that adaptation simply increases the size of the training set, or tracks systematic changes by exponentially weighting all previously seen data. This system was tested on the 35-word 10,710 token Lincoln Stressed Speech Data Base. Speaker adaptation experiments produced error rates equivalent to speaker-trained systems after the presentation of only a single new token per vocabulary word. Stress condition adaptation experiments produced results comparable to multistyle-trained systems after the presentation of several new tokens per vocabulary word.

Introduction

An isolated-word recognition system has been developed which dynamically adapts to changes in input speech. Two scenarios for adaptation were explored. In the first scenario the recognition system adapts to new previously unheard speakers. A system such as this could be used to address the problem of speaker independence, and could also provide a means for rapidly generating word models for a new speaker by using a previous speaker's word models as an initial state. Work in this area was previously presented by Baker in [1]. In the second scenario the recognition system adapts to variations in the input speech caused by changes in the user environment. These changes could include extraneous noise or variations induced by stress on the user. Applications of such a system are very appealing because of its ability to adapt to conditions which cannot be replicated during training, or which cannot be anticipated during recognition.

Both scenarios assume that some degree of user feedback is available to the system. In the most limited case the user must be able to reject incorrect recognition responses. This allows the adaptive system to use all correctly recognized tokens as additional data for estimating model parameters. A less limited system will allow more information to be fed back to the system. Such a system allows the user to correct errors made by the recognizer. This allows the adaptive system to make use of all input tokens regardless of whether they were correctly recognized or not.

The rate at which the model parameters must adapt to new input largely depends on the application of the system. In this system two methods of modifying adaptation rates were explored. The first method weights all available data equally regardless of when specific data was observed. This method has the effect of increasing the training set from which estimates are made. This feature can be very valuable to systems where limited training data is a factor. The second method weights more recently seen data more heavily than less recently seen data. The data is given an exponential weight which decays with time at some predetermined rate. This method enables the models to adapt quickly from one environment to another.

System Description

The adaptive recognition system used an isolated word system described by Paul in [2]. Each word was modeled by a 15-node Hidden Markov Model (HMM) which included two background nodes and a degenerate absorber node. The distributions for each parameter within a node were modeled as Gaussian. Means were estimated for each word model, node, and parameter. Grand variance estimates for each observation vector parameter were used for all variance calculations. A mel-cepstral-based front end was used to reduce a 20 ms frame of speech to 13 parameters. This analysis was done every 10 ms.

All experiments were performed on the Lincoln Stressed Speech Data Base [3]. The vocabulary consisted of 35 highly confused words and the entire data base contained over 10,710 tokens. Words were spoken by 9 speakers over a variety of talking styles and conditions, which included 7 simulated stress conditions and 3 real stress conditions, in addition to normal speech.

Initial word models for all experiments were obtained from 5 normally spoken tokens of each vocabulary word from a single speaker. The Baum-Welch (forward-backward) algorithm was used to estimate the HMM parameters for each word model.

During recognition, the Viterbi algorithm was used to decode all input tokens once for each of the vocabulary's word models. A backtrace of this decode was kept during recognition. From this, an assignment could be made for each observation frame to a HMM node. This has the effect of segmenting the input token. This segmentation was then used as part of the adaptation procedure to re-estimate the parameters of the model. Nodal means were estimated by averaging the observation vectors assigned to each node. Nodal transition probabilities were estimated based on the number of observations assigned to each node.

Experiments

All experiments performed were designed to investigate certain aspects relating to either of the two scenarios described above: speaker adaptation and stress condition adaptation.

Speaker Adaptation

All speaker adaptation experiments were performed on normal speech. Stressed speech was not used so as to separate the effects of speaker variation and stress condition variation. The Lincoln Stressed Speech Data Base contains a total of 13 normally spoken tokens for each of the 9 speakers and each of the 35 words in the vocabulary. Therefore, a maximum of 13 passes could be made through the entire vocabulary during a recognition test for a single speaker. Experiments were performed by selecting 9 speaker pairs where the models for the first speaker were adapted to the speech of the second speaker. Initial word models were generated from 5 normally spoken tokens of each vocabulary word from a single speaker. A recognition list was prepared which included 13 passes through the vocabulary as described above. During recognition, the first pass through the vocabulary was recognized using the initial word models. After each pass through the vocabulary was completed, all word models were adapted to incorporate the data from that pass. Results were then accumulated based on the number of errors found through each pass through the vocabulary. In this way the evolution of the word models as they adapted could be analyzed in terms of their recognition performance. Results shown in Figure 1 plot word error rate versus the number of new tokens observed for each word model. Error rates are shown for 5 cases: a non-adaptive system, a system that adapts only to correctly-recognized words, a system that adapts to all new input tokens, a system that adapts to all new tokens using an exponential weighting scheme, and a non-adaptive speaker-specific-trained system. The performance of the non-adaptive speaker-specific system represents the ultimate desired performance level for the adaptive systems. The performance of the non-adaptive system without the initial speaker-specific models represents the starting point for adaptive system improvements. While performance was not expected to surpass the speaker-specific system, the adaptive system was hoped to approach this level of performance. In Figure 1 the performance for the non-adaptive system is quite poor. The average error rate is above 20%. When adaptation was performed using only correctly-recognized tokens, the performance improved, but only slightly. When adaptation was performed using all input tokens, the improvement was dramatic. After only a few new tokens per word model, the error rates were reduced to a statistically insignificant difference from the speaker-specific-trained system. When an exponential weight was used in the adaptation, algorithm improvements occurred more quickly after only a single new token was presented for each vocabulary word.

The poor performance of the non-adaptive system was not unexpected. Speech variations from talker to talker are great enough to substantially degrade performance. When adaptation is used to include only correctly recognized tokens, improvements are quite modest. Any drastic changes, resulting from a new speaker, will very likely result in an error. This key information, therefore, is not used by the adaptive system in this feedback scenario. Because of this, only very minor modifications to the word-model parameters are possible when adaptation is done on

only correctly-recognized tokens. When adaptation is done on all input tokens, the most significant changes in the input speech can be captured by the adapting models. Very severe modifications to the word-model parameters are possible with this type of adaptation. The error rates shown in the plot fall quite quickly and approach the level of the speaker-specific-trained system after only a few new tokens per vocabulary word. When an exponential weight was used in the adaptation algorithm, the improvements found were very similar to those above. The weighting scheme used resulted in word models being based almost entirely on the previous 6 tokens seen for each word. Giving added weight to the most recently seen words enables the model parameters to adjust very quickly to the systematic changes found with a new speaker.

Stress Condition Adaptation

Similar experiments were performed to investigate the performance of adaptation with stressed speech. All of these experiments used training and test data from the same speaker. Initial word models were once again generated from 5 normally spoken tokens of each word. Excluding these 5 tokens, the Lincoln Stressed Speech Data Base contained 29 additional tokens per word for each speaker. These were the tokens that were used for testing. They contained both normal, stressed, and simulated stressed speech. Recognition lists were generated by randomizing the test tokens so that each group of 35 words, representing a pass through the vocabulary, would contain a mixture of the various stress and simulated stress conditions. Creating lists in this manner enabled a good representation of how the performance improved with increasing adaptation data for a range of stress conditions. Results of several recognition tests are shown in Figure 2. Recognition runs were performed on all 9 speakers in the data base and were averaged to produce the results shown. The figure plots error rate versus the number of tokens presented per vocabulary word. Points are plotted at increments of 5 new tokens per vocabulary word, to smooth the data. The results of a single pass through the vocabulary appeared to be very sensitive to the variations in the distribution of stress conditions contained in the pass, due to the randomization of the recognition lists. By smoothing over 5 passes, this variation was reduced. The 4 curves on this plot correspond to: a non-adaptive system, an adaptive system which uses only correctly recognized words, an adaptive system which uses all input tokens, and a speaker-specific multistyle-trained non-adaptive system. This multistyle training technique was first presented in [3].

The worst performance shown on this plot is markedly superior to the worst performance found in Figure 1. This suggests that variations due to a speaker change are much more severe (in terms of recognition performance) than the variations resulting from stress or simulated stress conditions. The non-adaptive system resulted in an average error rate of 3.4%. The multistyle-trained system resulted in an average of 1.6%. The results with the adaptive systems are similar to those presented in Figure 1, but are much less dramatic. The variations found with the stressed speech are not generally as systematic as the changes with a new speaker. The addition of new tokens can best be used, therefore, to effectively increase the training set size. The adaptation algorithm used equal weighting for all input tokens to best represent this. The conclusion is that with variations as wide as those found with stressed speech many new tokens are re-

quired to substantially reduce the error rate. However, the error rate can be reduced to levels that are comparable to multistyle-trained systems.

Conclusions

An adaptive recognition system has been developed which adjusts its model parameters to variations found with new speakers and various stress environments. Feedback was found to be necessary for implementing such a system. The facility in a system for the user to correct errors made by the recognizer, as opposed to simply rejecting, them was found to be essential for producing reasonable performance improvements over non-adaptive systems. Finally, in both the speaker adaptation and stress condition adaptation scenarios, the adaptive system resulted in error rates that were comparable to speaker-trained or multistyle-trained systems.

References

- [1] J. M. Baker and D. F. Pinto, "Optimal and Suboptimal Training Strategies for Automatic Speech Recognition in Noise and the Effects of Adaptation on Performance," ICASSP'86, Tokyo, Japan, April 1986.
- [2] D. B. Paul, "A Speaker Stress-Resistant HMM Isolated Word Recognizer," ICASSP'87, Dallas, TX, April 1987.
- [3] R. P. Lippmann, E. A. Martin, D. B. Paul, "Multi-style Training for Robust Isolated Word Speech Recognition," ICASSP'87, Dallas, TX, April 1987.

¹This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government

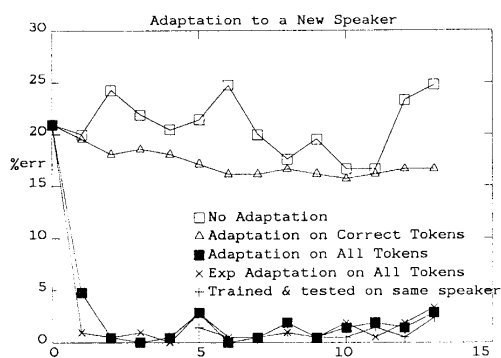


Figure 1: Error rates plotted as a function of the number of input tokens presented for each word. All points are averaged over 9 speaker pairs where initial training models are obtained from one speaker then tested on another.

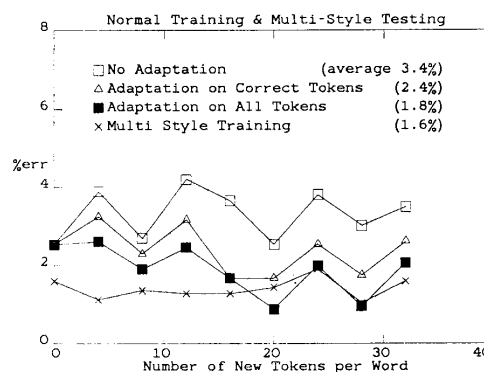


Figure 2: Error rates plotted as a function of the number of words presented for each vocabulary word. The plot has been smoothed by only showing averaged results for every 5 additional tokens. All tests were speaker-dependent and points shown are averaged over the 9 speakers in the data base.