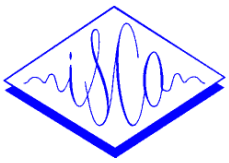
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

HYBRID HIDDEN MARKOV MODELS IN SPEECH RECOGNITION

Z. Li, P. Kenny and D. O'Shaughnessy

INRS-Télécommunications, Université du Québec
16 Place du Commerce, Verdun, Québec, Canada H3E 1H6

ABSTRACT

In speech recognition systems based on Hidden Markov Modeling, the computation of the likelihoods in detailed models is intensive, while the performance of crude models is poor. A hybrid model which combines the detailed and crude models is proposed to take advantage of the performance of the detailed model and the speed of the crude model. Experimental results show that a significant (up to a factor of 20) likelihood computation reduction has been obtained, with almost the same recognition accuracy as the baseline models on both the speaker-dependent and speaker-independent systems.

10.21437/Eurospeech.1995-182

1. INTRODUCTION

In this paper, we describe our efforts in acoustic modeling towards a very large-vocabulary continuous-speech recognizer. Our algorithm is a two-pass approach [1, 2], in which the first pass produces a word graph which should contain the correct hypothesis using crude acoustic and language models, and the second pass rescores the results of the first pass using detailed acoustic and language models in an effort to locate the correct hypothesis from the word graph.

In selecting the technique of acoustic modeling by using Hidden Markov Models (HMMs) in the first pass, the trade-off between speed and accuracy has to be considered. As we know, the continuous models have a superior performance in terms of accuracy, but they are very slow due to a huge amount of Gaussian distributions to be evaluated every frame. Although state-based clustering [3] and Gaussian clustering [4, 5] show an effective way to reduce the number of Gaussian distributions, but they still need to calculate far too many Gaussian distributions.

With tied-mixture models the distance computation is reduced, but the likelihood computation is still intensive, as discussed in [6]. An alternative would be to use discrete models, but the performance with discrete models is degraded.

In order to calculate the likelihoods efficiently, we propose a new type of model which we refer to as a *hybrid* HMM. The hybrid model essentially combines two different types of HMMs into one. If the hybrid model is the combination of a detailed model and a crude model, it can take advantage of the accuracy given by the detailed model and the speed given by the crude model. In the experimentation we further restrict the crude model to a discrete model, and choose a tied-mixture model as the detailed model.

2. HYBRID MODELS

In the first pass of our recognition system we firstly search a diphone graph by means of a backward Viterbi search to exhaustively draw up a table of estimates of phone scores and durations [7]. These estimates of phone scores and durations can be used to calculate an approximate acoustic match for an arbitrary phonetic transcription at a cost of a single floating point operation. To calculate these estimates

rapidly, simple models have to be used. Accordingly, we used tied-mixture models as baseline models in the first pass.

2.1 Tied-Mixture Models

We begin with the acoustic front-end. The speech signal is sampled at 16 kHz. A 30 ms Hanning window with 10 ms advance is used to compute the observation vectors based on Mel-frequency cepstral analysis. We compute 15 cepstral coefficients (static parameters) and append the vector with first difference cepstral coefficients (dynamic parameters) resulting in a 30 dimensional observation vector, once every 10 ms.

In the acoustic modeling, we use HMMs with a simple topology, namely models having only two or three states in addition to the sink state and no skip transitions. Output distributions are associated with states rather than transitions. Furthermore all transition probabilities are assumed to be the same. For modeling the output distributions, we used two codebooks, one for static parameters (except the power C_0) and another for dynamic parameters. For tied-mixture models, a single full-covariance matrix and a set of 256 means are used to tie all distributions.

In the first pass, we currently restrict ourselves to right-context models. Full right-context distributions are trained by using the Viterbi algorithm, then we cluster those distributions which have observations less than a minimum count M into context-independent distributions. In recognition, we use the right-context distributions with more than M training observations and the context-independent distributions. We treat M as a tuning parameter. After clustering the distributions, we smooth zero weights by backing off to a flat distribution with a penalty. A similar idea has been used in smoothing language models [8]. The penalty for the zero-count weights is given by the following formula

$$\frac{1}{\sum_i c[i] + 1}$$

where $c[i]$ is the observation count of the ith weight for a distribution, and the sum is over all weights of the distribution.

By using the tied-mixture models, our speaker-dependent system ended up about ten times slower than real-time. To speed up the likelihood computation of the tied mixture models, we tried to prune the 256 weights for each distribution, but the pruning was not effective. We also applied Gaussian shortlists [4, 5] to the tied-mixture models. Since we only have one covariance matrix and 256 means for all distributions, if we use vector quantization with one covariance matrix and 256 means to subdivide the acoustic space of the tied mixture models, it results in a quantizer with the same covariance matrix and the same means as in the tied-mixture models with zero quantization distortion. Therefore we may associate each codeword of the quantizer with the corresponding Gaussian (in fact there is only one Gaussian associated with a codeword with zero quantization distortion). In our system, the accuracy is degraded by using this kind of Gaussian shortlists.

2.2 Hybrid Models

The construction of the hybrid model is based on the following observation. The reason that the tied-mixture models outperform the discrete models is largely because the covariance matrices and means in the tied-mixture models are better estimated than in the discrete models. Therefore, we may use these better covariance matrices and means to produce a new covariance matrix and a new set of means, which can be used in a discrete manner. The clustering procedure can be done by using the quantization distortion criteria. The resulting Gaussians are then defined to be the Gaussians of the hybrid models. The distributions of the hybrid models remain undefined at this moment.

Recall that in the Gaussian shortlists approach a vector quantizer is only used as an intermediate step to obtain a codeword which associates to a list of Gaussians. So there is no need to define distributions in the vector quantizer. Since we want to use the hybrid models like discrete models, the weights of the distributions have to be provided. We cannot just copy the weights of distributions in the tied-mixture models since the reestimation formula for the weights in the discrete

models is different from that in the tied-mixture models. Accordingly, it is necessary to do a single iteration of the reestimation of the weights of the hybrid models in a discrete manner. In summary, the hybrid models take the Gaussians from clustering the Gaussians of the tied-mixture models and reestimate weights of the distributions as in the discrete models. These hybrid models are then the hybrids of tied-mixture models and discrete models.

To train the hybrid models, we firstly train the tied-mixture models. As mentioned in the last subsection, we trained full-right context distributions with one covariance matrix and 256 means. In the hybrid models, we also use one covariance matrix and 256 means, so the clustering of the Gaussians reduces to just copying the Gaussians of the tied-mixture models with zero quantization distortion. Then we reestimate the weights of distributions in a discrete manner by a single iteration of reestimation. In recognition, we use the same method as in the tied-mixture models for clustering distributions and smoothing zero weights. To calculate the likelihoods, the hybrid models are treated as discrete models, so it is very fast.

3. EXPERIMENTAL RESULTS

The purpose of the second pass is to use expensive models to rescore the word graph. However, in order to check the performance of the first pass acoustic models, we may use these models in the second pass as well. In all the following experiments, we used the same acoustic models in both passes. To test the hybrid models, we firstly applied them to the speaker-dependent continuous-speech recognition system. Recently the hybrid models were also developed for the speaker-independent system. With the hybrid models, we have achieved a speed as fast as the discrete models, which is about 20 times faster than the tied-mixture models in calculating the likelihoods. As for the recognition accuracy, we have compared the hybrid models with the tied-mixture models on both the speaker-dependent system and the speaker-independent system.

3.1 Speaker-Dependent System

The speaker-dependent system is trained with 3 hours of acoustic training data collected from a single speaker. It uses a 4,900-word vocabulary. A bigram language model with perplexity 143 is used for both the first and second passes. The parameters of the recognizer were tuned on a 556-word development file and tested on 2290-word evaluation files. The acoustic models are HMMs with two states and two codebooks, one for 14 static cepstral parameters and another for 15 dynamic cepstral parameters. We first trained the tied-mixture models with one covariance matrix, 256 means and full right-context distributions. Then the hybrid models were developed. The comparison between the tied-mixture models and the hybrid models is given in Table 1. The word error rate is 8.69% for the tied-mixture models and 8.03% for the corresponding hybrid models in the evaluation set. The system (including the acoustic front-end) with hybrid models runs real-time on an HP 735 work station.

	Ins %	Del %	Sub %	Word Error %
tied-mixture	1.27	1.79	5.63	8.69
hybrid	1.44	1.75	4.85	8.03

Table 1. *Comparison of modeling using tied-mixture models versus hybrid models in the speaker-dependent system.*

3.1 Speaker-Independent System

In developing the system, we have used the ARPA *Wall Street Journal*-based speaker independent CSR corpus. We have trained our acoustic models by using WSJ0 data, and had separate male and female models. We used acoustic models with three-state HMMs in the speaker-independent system rather than two-state HMMs in the speaker-dependent system. The other set-up of the acoustic modeling is the same as

in the speaker-dependent system. We firstly trained tied-mixture models, and then hybrid models. We have tested both the tied-mixture models and the hybrid models on WSJ0 development set for four female speakers. The vocabulary size is 5,000 words. In the first pass a bigram language model with perplexity 128 was used to conduct the search. In the second pass a trigram language model with perplexity 104 was incorporated. Table 2 gives the comparison of the word error rates between the tied-mixture models and the hybrid models for four speakers. The average word error rate of the four speakers is 12.65% for the tied-mixture models and 11.76% for the hybrid models.

Speakers	050	053	420	421	average
tied-mixture	7.49	14.73	11.87	16.57	12.65
hybrid	7.04	16.79	9.84	13.37	11.76

Table 2. *Comparison of modeling using tied-mixture models versus hybrid models in word error rate (%) in the speaker-independent system.*

4. CONCLUSIONS

The hybrid model which combines the detailed and crude models was proposed to take advantage of the superior performance of the detailed model and the superior speed of the crude model. Particularly, the hybrid models from combinations of tied-mixture models and discrete models were developed for both the speaker-dependent and speaker-independent systems. Significant (up to a factor of 20) likelihood computation reductions have been achieved on both systems, with almost the same recognition accuracy as our baseline tied-mixture models.

In principle, hybrid models could be the combination of any detailed models and crude models, such as choosing the detailed models as continuous models and crude models as discrete models. In order to have a reasonable resolution, we could increase the codebook size in the hybrid models and then speed up the distance calculation by a tree search.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] P. Kenny, P. Labute, Z. Li and D. O'Shaughnessy, "New graph search techniques for speech recognition," *Proceedings ICASSP 94*, vol. 1, pp. 553–556, April 1994.

[2] Z. Li , P. Kenny and D. O'Shaughnessy, "Searching with a transcription graph," *Proceedings ICASSP 95*, vol. 1, pp. 564–567, May 1995.

[3] M.Y. Hwang and X.D. Huang, "Subphonetic modeling with Markov states – senone," *Proceedings ICASSP 92*, vol. 1, pp. 33–36, March 1992.

[4] E. Bocchieri, "Vector quantization for the efficient computation of continuous density likelihoods," *Proceedings ICASSP 93*, vol. 2, pp. 692–695, April 1993.

[5] V. Digalakis and H. Murveit, "Genones: optimizing the degree of mixture tying in a large vocabulary Hidden Markov Models based speech recognizer," *Proceedings ICASSP 94*, vol. 1, pp. 537–540, April 1994.

[6] D.B. Paul, "The Lincoln tied-mixture HMM continuous speech recognizer," *Proceedings ICASSP 91*, pp. 329–332, May 1991.

[7] P. Kenny, P. Labute, Z. Li, R. Hollan, M. Lennig and D. O'Shaughnessy, "A very fast method for scoring phonetic transcriptions," *Proceedings Eurospeech 93*, vol. 3, pp. 2117–2120, September 1993.

[8] P. Placeway, R. Schwartz. P. Fung and L. Nguyen, "The estimation of powerful language models from small and large corpora," *Proceedings ICASSP 93*, vol. 2, pp. 33–36, April 1993.