Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition

Xuedong Huang, Fil Alleva, Satoru Hayamizu
Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

This paper reports recent efforts to further improve the performance of the Sphinx system for speaker-independent continuous speech recognition. The recognition error rate is significantly reduced with incorporation of additional dynamic features, semi-continuous hidden Markov models, and speaker clustering. For the June 1990 (RM2) evaluation test set, the error rates of our current system are 4.3% and 19.9% for word-pair grammar and no grammar respectively.

Introduction

This paper reports our recent effort to further improve the accuracy of the Sphinx System [10]. We choose here to adhere to the basic architecture of the Sphinx System, and to use the standard Resource Management task and training corpus. Possible improvements could be resulted from the following categories:

- Incorporate additional dynamic features.
- Improve HMM probabilistic representation.
- Cluster training speakers to provide multiple models.
- Introduce discriminant transformations to improve discrimination.
- Extend corrective training for semi-continuous models.
- Improve allophonetic models of coarticulation.

This paper will report advances on the first five categories. Improved allophonic modeling has been reported in [13].

Our first improvement involves the incorporation of dynamic features computed from the LPC cepstrum. Previous versions of the Sphinx system have used first order differenced cepstrum and power. Here, we experimented with second and third order differenced cepstrum and power. We also experimented with incorporation of both 40 msec and 80 msec differenced cepstrum, as well as the difference derived from compressed speech [3]. These additional feature sets are incorporated in the multi-codebook framework. The best combination reduced errors by 17% over the baseline results.

Our second technique employs the semi-continuous hidden Markov model (HMM) [8]. Multiple-codebook semi-continuous models are extended to our current Sphinx version. Both diagonal and full covariance Gaussian models are investigated. We found that the best variants of both models reduced error rate of discrete HMMs by 10-20%.

Due to smoothing abilities of the semi-continuous model, we were able to train multiple sets of models for different speakers. We investigated automatic speaker clustering and explicit male/female clustered models. In both cases, models of all the speaker clusters are simultaneously active, with the restriction that no between-cluster transitions are allowed. Thus, the system retains speaker-independent characteristics. By using multiple model sets with the semi-continuous HMM, the error rate is further reduced by 10-15%.

We experimented with two variants of linear discriminant transformations. The first attempted to use a single transformation to separate all triphone states. The second attempted to shift the mean vectors of the semi-continuous mixtures, so as to separate confusable words. However, neither method produced any improvement.

Finally, we investigated corrective training for semi-continuous models. At the time of this writing, we have only applied our discrete corrective training algorithm [15] to semi-continuous models. We found that this method is effective if top-1 (or discrete HMM) decoding is used. However, if the recognition algorithm considers top N codewords, while the corrective training uses only the top 1 codeword, the results degrade considerably. Thus, corrective training is not used in this evaluation.

In the next five sections, we describe these techniques. We will measure improvements based on our baseline system as reported in [11], and evaluated on the 600 sentences that comprise the February and October 1989 test sets. Next, a summary of all the improvements will be provided for the tuning (February and October 1989) sets, as well as the new RM2 test set (480 sentences from 4 speakers). The last section contains our conclusion and outlines our future work.

Dynamic Features

Temporal changes in the spectra are believed to play an important role in human perception. One way to capture this information is to use delta coefficients or differenced coefficients [4, 14] that measure the change of coefficients over time. Temporal information is particularly suitable for HMMs, since HMMs assume each frame is independent of the past, and these dynamic features broaden the scope of a frame.

In the past, the Sphinx system has utilized three codebooks containing:

- 12 LPC cepstrum coefficients.
- 12 differenced LPC cepstrum coefficients (40 msec. difference)
- Power and differenced power (40 msec.).

We experimented with a number of new measures of spectral dynamics, including:

- Second order differencing (cepstrum and power).
- Third order differencing (cepstrum and power).
- Multiple window differencing (40 msec. and 80 msec).
- Differencing from temporally compressed speech.

The first set of coefficients is incorporated in a new codebook, whose parameters are second order differences of the cepstrum coefficient. The second order difference for frame n is the difference between n+1 and n-1 first order differential coefficients. We incorporated this as a fourth codebook, and evaluated the new system using the word pair grammar (perplexity 60) on the February and October 1989 test sets (600 sentences). We found that second order differencing reduced errors from 6.9% to 6.2%. Second order differencing of power (used as another parameter in the power codebook) further reduced errors to 6.0%.

We attempted to extend this idea to third-order differencing, taking the difference of adjacent second-order differential coefficients. But we found that performance deteriorated slightly. We conclude that there is little information beyond second-order differences.

Next, we incorporated both 40 msec. and 80 msec. differences, which represent short-term and long-term spectral dynamics, respectively. We hoped that these two sources of information are more complementary than redundant. We first incorporated the two as separate codebooks (making a total of five codebooks), which reduced errors from 6.0% to 5.9%. We then incorporated the two into one codebook, weighted by their variances. This further reduced errors to 5.7%. We believe the latter approach gave better performance because the correlation between the 40 msec. and 80 msec. differences violated the codebook independence assumption of the multi-codebook approach.

Finally, we tried to incorporate a variable measure of spectral dynamics. Instead of taking static differences, we take differences that depend on "acoustic segments." We defined acoustic segments by using the variable frame rate method [16]. Speech is segmented according to the Euclidean distance of the cepstral coefficients. A segment boundary is placed between frames whose distance exceeds a pre-set threshold. The threshold is chosen so that the ratio of frames to segments is about 2.5 to 1. Each segment is then averaged into a single cepstral (and power) vector. The differential measure for segment n is computed by subtracting the averaged cepstrum of segment n-1 from that of n+1. Then, the compressed cepstrum is expanded back to its original frame length, by duplicating the compressed frames, so that its length matches that of the other code sequences. This provides more acoustic context for frames that are in stationary regions. We used this codebook instead of the second order differences, and found that errors increased to over 7%. One explanation for this phenomenon is that this type of compression-expansion increased frame-to-frame correlation, which makes HMMs less appropriate models.

Thus, the final configuration involves four codebooks, each with 256 entries, that use:

- 12 LPC cepstrum coefficients.
- 12 40-msec differenced LPC cepstrum coefficients and 12 80-msec differenced LPC cepstrum coefficients.
- 12 second-order differenced power.
- Power, 40-msec differenced power, second-order differenced power.

This configuration reduced an original error rate of 6.9% to 5.7%, a 17% error rate reduction. A summary of dynamic feature results is give in Table 1.

Systems	Error Rate
Baseline	6.9%
Additional dynamic features	5.7%

Table 1: Improvements using additional dynamic features.

Semi-Continuous HMMs

Semi-continuous hidden Markov models mutually optimize the VQ codebook and HMM parameters under a unified probabilistic framework [7, 8, 6]. Here, each VQ codeword is regarded as a continuous probability density function. Intuitively, from the discrete HMM point of view, semi-continuous HMMs integrate quantization accuracy into the HMM, and robustly estimate the discrete output probabilities by considering multiple codeword candidates in VQ procedure. From the continuous mixture HMM point of view, semi-continuous HMMs employ a shared mixture of continuous output probability densities for each individual HMM. Shared mixtures substantially reduce the number of free parameters and computational complexity in comparison with the continuous mixture HMM, while maintaining reasonably its modeling power. For the semi-continuous model, appropriate acoustic representation and probability density functions is crucial to the recognition accuracy. With appropriately chosen acoustic parameters and probability density functions, the semi-continuous HMM can greatly enhance the robustness in comparison with the discrete HMM [8].

We first performed exploratory semi-continuous experiments on our three-codebook system. The semi-continuous HMM was extended to accommodate multiple feature front-end [8, 6]. All codebook means and covariance matrices are reestimated together with the HMM parameters except the power covariance matrices, which are fixed. In an early experiment on the June 88 test set, we found that full covariance HMMs outperformed diagonal covariance semi-continuous HMMs (with an error reduction of 10% in comparison with the diagonal semi-continuous models, and 20% error reduction in comparison with the discrete HMM). However, on the present tuning set, the full covariance semi-continuous HMMs did not give us any improvement. This is probably because the correlation among our acoustic parameters is not very strong, so that the diagonal covariance assumption is relatively valid. When three codebooks are used, the diagonal semi-continuous model reduced error rate of the discrete HMM by 13%. Results using three codebooks are shown in Table 2.

Models	Error Rate
Discrete HMM	6.9%
Semi-continuous HMM	6.0%

Table 2: Discrete and semi-continuous results for three codebook systems.

Another advantage to use the semi-continuous HMM is that it requires less training data in comparison with the discrete HMM. Therefore, given current training data set, more detailed models can be employed to improve the recognition accuracy. One way to increase the number of parameters is to use speaker-clustered models as shown in the following section.

Speaker Clustering

In the past, we have experimented with speaker clustering as a means of speaker adaptation [12]; however, we found that clustering fragmented the training data, and actually degraded performance. In that experiment, no smoothing across cluster was performed. We now rectify this problem with two different approaches.

The first approach uses discrete models, and smoothes them using deleted interpolation between correct cluster and other clusters. We clustered the speakers based on similarity of their allophonic HMMs [5]. To perform recognition, one recognition network is generated for each speaker cluster. All networks are run in parallel, and the best overall scoring path is chosen as the recognized sentence. Note that this is a speaker-independent approach, as no a priori cluster selection takes place. With two and three clusters, this approach reduced errors by about 6%.

The second approach smoothes the resulting models by semi-continuous HMMs. Because multi-codewords are used in Forward-Backward training for semi-continuous models, more models can be trained robustly. Thus, smoothing takes place only within-cluster, and not between-cluster. For this study, we simply used male and female as the two clusters. No interpolation between clustered models is used. The best overall scoring path with clustered models is chosen as the recognized sentence. For three-codebook systems, the error reduction of clustered semi-continuous HMMs is over 10% in comparison with the semi-continuous HMM, and over 20% in comparison with the clustered discrete HMM.

Finally, we combined the four-codebook front-end with the speaker-clustered semi-continuous HMMs. The results are shown in Table 3. The combined error reduction here is 17% in comparison with the discrete HMM.

Systems	Error Rate
Discrete HMM	5.7%
Semi-continuous HMM	4.7%

Table 3: Four codebook results: discrete HMMs vs. speaker-clustered, semi-continuous HMMs.

Discriminant Transformations

Two variants of linear discriminant transformation were experimented. First, the classes to be discriminated are defined as triphone states. The Viterbi segmented data are used to compute within- and between-class means and covariance matrices. Here, 7 continuous frames are treated as one vector for discriminate transformation. The transformed vector corresponding to top three-frame eigenvalues are divided into three vectors for three-codebook generation. Several variations of the approach were experimented. However, the average recognition accuracy is not improved.

Next, we experimented with a unified linear discriminant transformation to find appropriate features for semi-continuous hidden Markov modeling. We used word level supervision to estimate the confusion covariance matrices. This extends the technique suggested by [9, 2] to the semi-continuous HMM. Both within- and confusion-covariance matrices for each VQ codeword are weighted with the semi-continuous HMM posterior probabilities. We investigated both codeword-dependent and codeword-independent discriminant transformations with different parameters. Unfortunately, the final word accuracy is still about the same as our best semi-continuous HMM.

Results of the unified discriminat transformation were promising. We think more experiments are needed to fully understand the problem.

Corrective Training

Previously, we have applied the IBM corrective training algorithm [1] to continuous speech training [15]. This approach basically involved generation of misrecognitions and near-misses for each training sentence, and then modifying the HMM parameters to discriminate the correct sentence from the misrecognitions and near-misses.

For discrete models, this method rewards codewords that contribute to the correct alignment, and punishes those that contribute to misrecognitions and near-misses. However, with a semi-continuous system, several codewords are accountable for each frame alignment. At the time of this writing, we have only used a simple extension of our algorithm: for the purpose of corrective training, only the top semi-continuous candidate (rather than top 4 or 6) was used.

This technique essentially uses top-1 correction and top-4 decoding. We found that this technique increased errors substantially, presumably due to the mismatch between the corrective and decoding stages. In a second experiment, both top-1 correction and decoding were applied (although hypotheses were generated with a top-4 system), significant improvements were observed (an error reduction of 10-15%). However, the improvement was less than that of the 4-codebook semi-continuous HMM. Thus, for evaluation purposes, we opted to bypass the corrective training stage.

In order to reap maximum benefit from corrective training, we will need to implement a consistent algorithm for semi-continuous corrective training. We also believe that an N-best algorithm [17] for hypothesizing near-misses will help significantly.

Summary of Results

Without corrective training, our previous best results was 6.9% error rate on the 600 sentence tuning set (with corrective training, this was reduced to 5.7%). We will refer to the 6.9% error rate system as the "baseline" system. Table 4 shows our progress with the techniques described in this paper. This represented a 32% error rate reduction from the baseline system. We believe with proper implementation of corrective training, another 10% or more reduction will be possible.

Systems	Error Rate
Baseline	6.9%
+2nd order diff. cepstrum	6.2%
+2nd order diff. power	6.0%
+80ms 1st diff. order cepstrum	5.7%
+Semi-continuous clustered model	4.7%

Table 4: Improvements of various techniques using the word-pair grammar.

Since our intermediate results were only evaluated on the word-pair system, we do not have detailed results for the no-grammar system. The baseline and final system results are shown in Table 5. The improvements introduced here led to a 28% error reduction.

Finally, we evaluated the above system on the June 90 (RM2) test set, which consists of 480 sentences spoken by four speakers. The evaluation results are shown in Table 6.

Systems	Error Rate
Baseline	27.1%
Final system	19.5%

Table 5: Improvements using no grammar.

The results on these speakers are better than the tuning set. The error reduction of our current system is about 40% in comparison with the baseline system. We believe this can be partially be attributed to the better modeling of female speech. Previously, speaker-independent models were trained with 1/3 female speech. With separated male/female models, female results improved substantially.

Speaker	Word-Pair Grammar Error Rate	No Grammar Error Rate
BJW	3.1%	18.6%
JLS	4.8%	21.3%
JRM	5.8%	24.0%
LPN	3.6%	15.7%
Average	4.3%	19.9%

Table 6: Results with RM2 test set.

Conclusions

In this paper, we have presented several techniques that substantially reduced Sphinx's error rate. These techniques include: dynamic features, semi-continuous HMMs, and speaker clustering. We have also found that discriminant transformations and dynamic features based on variable frame analysis did not improve recognition. We also obtained disappointing results using a compromised corrective training algorithm.

In the future, we expect to further extend some of these areas. We will investigate other methods for automatical parameter selection. We will extend speaker clustering to a much larger number of clusters (on a larger database). Corrective training could be improved by using N-Best sentence hypotheses, as well as by using a consistent algorithm for semi-continuous learning. Finally, we hope to further investigate discriminant methods, and learn whether they are limited to small vocabularies, or discover new variations that improve our large-vocabulary system.

We believe the improvement of basic speech research is essential for further progress of the spoken language systems. We hope extensions of the above areas of research will further narrow the gap of man-machine communication.

References

[1] Bahl, L., Brown, P., De Souza, P., and Mercer, R. *A New Algorithm for the Estimation of Hidden Markov Model Parameters.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing.** 1988.

[2] Doddington, G. *Phonetically Sensitive Discriminants for Improved Speech Recognition.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing.** 1989.

[3] Furui, S. *On the Use of Hierarchical Spectral Dynamics in Speech Recognition.* in: **IEEE International Con-**

ference on Acoustics, Speech, and Signal Processing. 1990, pp. 789–792.

[4] Furui, S. *Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum.* **IEEE Transactions on Acoustics, Speech, and Signal Processing**, vol. ASSP-34 (1986), pp. 52–59.

[5] Hayamizu, S., Lee, K., and Hon, H. *Description of Acoustic Variations by Hidden Markov Models with Tree Structure.* in: **International Conference on Spoken Language Processing**. 1990.

[6] Huang, X., Ariki, Y., and Jack, M. **Hidden Markov Models for Speech Recognition.** Edinburgh University Press, Edinburgh, U.K., 1990.

[7] Huang, X. and Jack, M. *Semi-Continuous Hidden Markov Models for Speech Signals.* **Computer Speech and Language**, vol. 3 (1989), pp. 239–252.

[8] Huang, X., Lee, K., and Hon, H. *On Semi-Continuous Hidden Markov Modeling.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing.** 1990.

[9] Hunt, M. *A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1989.

[10] Lee, K. **Automatic Speech Recognition: The Development of the SPHINX System.** Kluwer Academic Publishers, Boston, 1989.

[11] Lee, K. *Hidden Markov Models : Past, Present, and Future.* in: **Proceedings of Eurospeech**. 1989.

[12] Lee, K. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* Computer Science Department, Carnegie Mellon University, April 1988.

[13] Lee, K., Hayamizu, S., Hon, H., Huang, C., Swartz, J., and Weide, R. *Allophone Clustering for Continuous Speech Recognition.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing.** 1990.

[14] Lee, K., H.W., H., and Reddy, R. *An Overview of the SPHINX Speech Recognition System.* **IEEE Transactions on Acoustics, Speech, and Signal Processing**, January 1990.

[15] Lee, K. and Mahajan, S. *Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition.* **Computer Speech and Language**, vol. (1990), p. .

[16] Russell, M., Ponting, K., Peeling, S., Browning, S., Bridle, J., Moore, R., Galiano, I., and Howell, P. *The ARM Continuous Speech Recognition System.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1990, pp. 69–72.

[17] Schwartz, R. and Chow, Y. *The Optimal N-Best Algorithm: An Efficient Procedure for Finding Multiple Sentence Hypotheses.* in: **IEEE International Conference on Acoustics, Speech, and Signal Processing.** 1990.