On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition

X.D. Huang and K.F. Lee[1]

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

Speaker-independent system is desirable in many applications where speaker-specific data do not exist. Alternatively, if speaker-dependent data are available, the system could be adapted to the specific speaker such that the error rate could be significantly reduced. In this paper, we used DARPA Resource Management task as our domain to investigate the performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. We already have a state-of-the-art speaker-independent speech recognition system, SPHINX. The error rate for the RM2 test set is 4.3%. We extended SPHINX to speaker-dependent speech recognition. The error rate is reduced to 1.4 - 2.6% with 600-2400 training sentences for each speaker, which demonstrated a substantial difference between speaker-dependent and -independent systems. Based on speaker-independent models, we studied speaker-adaptive speech recognition. With 40 adaptation sentences for each speaker, the error rate can be reduced from 4.3% to 3.1%. When the number of speaker adaptation sentences is comparable to that of speaker-dependent training, speaker-adaptive recognition works better than speaker-dependent recognition, which indicates the robustness of speaker-adaptive speech recognition.

Introduction

Speaker-independent speech recognition systems could provide users with a ready-to-use system [14, 17, 10, 11]. We do not need to collect speaker-specific data to train the system, but collect data from a variety of speakers to reliably model many different speakers. Speaker-independent systems are definitely desirable in many applications where speaker-specific data do not exist. Alternatively, if speaker-dependent data are available, the system could be adapted to a specific speaker to further reduce the error rate. The problem of speaker-dependent systems is that for large-vocabulary continuous speech recognition, we generally need half an hour of speech from the specific speaker to reliably estimate the system parameters. The problem of speaker-independent systems is that the error rate of speaker-independent speech recognition systems is generally two to three times higher than that of speaker-dependent speech recognition systems [17, 10]. A logical compromise for a practical system is to start with a speaker-independent system, and then adapt the system to each individual user.

In this paper, we used DARPA Resource Management task as our domain to investigate the performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. We already have a state-of-the-art speaker-independent speech recognition system, SPHINX [14, 7]. The error rate for the RM2 test set, including two male and two female speakers with 120 sentences for each, is 4.3%. We extended SPHINX to speaker-dependent speech recognition, with 600 to 2400 speaker-dependent training sentences, the error rate is reduced to 2.6 - 1.4%. Not surprisingly, the error rate is reduced by two to three times in comparison with the speaker-independent system.

To bridge the gap between speaker-dependent and speaker-independent speech recognition, we want to modify two most important parameter sets for each speaker, i.e. the codebook mean vectors and the output distributions. Since adaptation is based on the speaker-independent system with only limited adaptation data, a good adaptation algorithm should be consistent with the speaker-independent parameter estimation criterion, and adapt those parameters that are less sensitive to the limited training data.

The codebook mean vector can represent the essential characteristics of different speakers, and can be rapidly estimated with only limited training data. Because of this, we consider it to be the most important parameter set. The semi-continuous hidden Markov model (SCHMM) [5] provide us with a good tool to modify the codebook for each speaker. With robust speaker-independent models, we modified the codebook according to the SCHMM structure such that the SCHMM likelihood can be maximized for the given speaker. This estimation procedure considers both phonetic and acoustic information, and gives us significant improvement when limited adaptation data are used.

Another important parameter set is the output distribution of the SCHMM. Since there are too many parameters in the output distributions, direct use of the SCHMM would not lead to any improvement. Analogous to Bayesian learning, we interpolate speaker-dependent output distribution with speaker-independent estimates. Due to limited adaptive data, we also measured the similarity between output distributions of different phonetic models, and grouped different distributions into several clusters. Finally, interpolation is carried out between original speaker-independent models and speaker-dependent clustered models.

With 40 adaptation sentences for each speaker, the above adaptation algorithms reduced the error rate to 3.1%. This error reduction is more than 25% in comparison with the best speaker-independent system on the same test set. Since our proposed algorithm can be used to incrementally adapt the speaker-independent system, we incrementally increased the adaptation sentences to 300-600, the error rate is further reduced with the increase of the adaptation data. With only 300 adaptation sentences, the error rate is lower than that of the best speaker-dependent system on the same test set (trained with 600 sentences).

Speaker-Independent System

Large-vocabulary speaker-independent continuous speech recognition has made significant progress during the past years [14, 17, 10, 11]. Sphinx, a state-of-the-art speaker-independent speech recognition system developed at CMU [14], has achieved high word recognition accuracy with the introduction and usage of the following techniques: (1) *multiple VQ codebooks*. In order to incorporate the multiple knowledge sources and minimize VQ errors, multiple vector quantized codebooks incorporating LPC cepstrum, differential cepstrum, second order differential cepstrum, and log-power parameters were used to model speech; (2) *generalized triphone models*. Triphones have been successfully used by [20, 18]. However, we believe that

[1] Dr. K.F. Lee is currently with Apple Computer, Inc.

CH2977-7/91/0000-0877 $1.00 © 1991 IEEE

many contexts are quite similar, and can be combined. Clustering contexts leads to fewer, and thus more trainable, models. We used an information-theoretic clustering procedure that combined similar contexts into generalized triphones [13]; (3) *between-word coarticulation modeling*. The concept of triphone modeling was extended to the word boundary, which leads to between-word triphone models [8]; (4) *semi-continuous models*. Semi-continuous hidden Markov models mutually optimize the VQ codebook and HMM parameters under a unified probabilistic framework [4]. Shared mixtures substantially reduces the number of free parameters and computational complexity in comparison with the continuous mixture HMM, while maintaining reasonably its modeling power. The SCHMM can greatly enhance the robustness in comparison with the discrete HMM [6, 7]; (5) *speaker-clustered models*. Another advantage to use the SCHMM is that it requires less training data in comparison with the discrete HMM. Therefore, given current training data set, speaker-clustered models (male/female in this study) were employed to improve the recognition accuracy [7].

Speaker	3990 Training Sent Word-Pair Grammar Error Rate
BJW	3.1%
JLS	4.8%
JRM	5.8%
LPN	3.6%
Average	4.3%

Table 1: Speaker-independent results with RM2 test set.

We evaluated the above system (without shared distribution modeling [7]) on the June 90 (RM2) test set, which consists of 480 sentences spoken by four speakers. The evaluation results are shown in Table 1. This will be referred as the baseline system in comparison with both speaker-dependent and speaker-adaptive systems.

Speaker-Dependent System

For speaker-dependent speech recognition, the training set consists of 600 sentences from each speaker. We used essentially the same system designed for the speaker-independent speech recognition. The SCHMM parameters and VQ codebook are estimated jointly starting with speaker-independent models. Two to three iterations were run to sharpen the output distributions. Results are listed in Table 2.

The average error rate for four speakers was reduced from 4.3% to 2.6%. This 40% error reduction demonstrated the importance of speaker-dependent data. When we further increased the training data of each speaker to 2400 sentences for each speaker, the error rate was reduced from 2.6% to 1.4%. The error rate of the speaker-independent system is about three times that of the speaker-dependent system, albeit this comparison is not fair since the speaker-independent system is trained with only 3990 sentences from about 100 speakers. However, these results clearly indicate the importance of training data. If speaker-dependent data are available, the error rate can be significantly reduced. Results presented here also demonstrated that techniques developed for speaker-independent speech recognition could be extended easily for speaker-dependent speech recognition.

Speaker-Adapative System

While last section clearly demonstrated the importance of speaker-dependent data, it is generally impractical for a speaker to speak 2400 sentences just for training. We are interested in using only a small amount of speaker-dependent data to adapt the speaker-independent models, so that an initially speaker-independent system can be rapidly improved as a speaker uses the system. We will examine how to adapt two most important parameter sets: the mean vector in the codebook and the output probability distributions in the SCHMM.

Speaker	600 Training Sent Error Rate	2400 Training Sent Error Rate
BJW	1.6%	1.0%
JLS	4.4%	2.7%
JRM	2.3%	1.5%
LPN	2.1%	0.4%
Average	2.6%	1.4%

Table 2: Speaker-dependent results with RM2 test set.

Codebook adaptation

The semi-continuous hidden Markov model (SCHMM) has been proposed to extend the discrete HMM by replacing discrete output probability distributions with a combination of the original discrete output probability distributions and continuous pdf of a codebook [5, 4]. In comparison with the conventional codebook adaptation techniques [21, 16, 15], the SCHMM can jointly reestimate both the codebook and HMM parameters in order to achieve an optimal codebook/model combination according to the maximum likelihood criterion. The SCHMM can thus be readily applied to speaker-adaptive speech recognition by reestimating the codebook.

With robust speaker-independent models, the codebook is modified according to the SCHMM structure such that the SCHMM likelihood can be maximized for a given speaker. Here, both phonetic and acoustic information are considered in the codebook mapping procedure since $Pr(X|M)$, the probability of acoustic observations X given the model M, is directly maximized. To elaborate, we first compute the posterior probability $\lambda_i(t)$ based on the speaker-independent model [4]. $\lambda_i(t)$ measures the similarity that acoustic vector at time t will be quantized with codeword i. The ith mean vector μ_i of the codebook can then be computed with

$$\mu_i = \frac{\sum_t \lambda_i(t) X_t}{\sum_t \lambda_i(t)} \tag{1}$$

Both the mean and variance vector can be adapted iteratively with Equation 1. However, the variances cannot be reliably estimated with limited adaptive data. Because of this, we can interpolate estimates with speaker-independent estimates analogous to Bayesian adaptation [2, 22]. However, in comparison with iterative SCHMM codebook reestimation, we have not achieved any significant error reduction by combining interpolation into the codebook mapping procedure. It seems to be sufficient by just using very few samples to reestimate the mean vector. Speaker-adaptive recognition results with 5 to 150 adaptive sentences from each speaker are listed in Table 3. Detailed results for 40 adaptive sentences are listed in Table 4.

In this study, we used the SCHMM to reestimate the mean vector only. Three iterations are carried out for each speaker. The error rates with 5 to 40 adaptive sentences from each speaker are 3.8% and 3.6%, respectively. In comparison with the speaker-independent model, the error rate of adaptive systems is reduced by about 15% with only 40 sentences from each speaker. Further increase in the number of adaptive sentences did not lead to any significant improvement.

Systems	Word Pair Grammar Error
Without adapt	4.3%
5 adapt-sent	3.8%
40 adapt-sent	3.6%
150 adapt-sent	3.5%

Table 3: Adaptation results with the SCHMM.

Output distribution adaptation

Several output-distribution adaptation techniques, including cooccurence mapping [12, 3], deleted interpolation [9, 4], and state-level-distribution clustering, are studied. All these studies are based on SCHMM-adapted codebook as discussed above.

In cooccurence mapping, the cooccurence matrix, the probability of codewords of the target speaker given the codeword of speaker-independent models, is first computed [3]. The output distribution of the speaker-independent models is then projected according to the cooccurence matrix. We did not obtain any improvement with cooccurence mapping. This is probably because that cooccurence smoothing only plays the role of smoothing, which is not directly related to maximum likelihood estimation.

A better adaptation technique should be consistent with the criterion used in the speech recognition system. As the total number of distribution parameters is much larger than the codebook parameters, direct reestimation based on the SCHMM will not lead to any improvement. To alleviate the parameter problem, the similarity between output distributions of different phonetic models is measured. If two distributions are similar, they are grouped into the same cluster in a similar manner as the generalized triphone [12]. Since clustering is carried out at the state-level, it is more flexible and more reliable in comparison with model-level clustering. Given two distributions, $b_i(O_k)$ and $b_j(O_k)$, the similarity between $b_i(O_k)$ and $b_j(O_k)$ is measured by

$$d(b_i . b_j) = \frac{(\prod_k b_i(O_k)^{C_i(O_k)})(\prod_k b_j(O_k)^{C_j(O_k)})}{(\prod_i b_{i+j}(O_k)^{C_{i+j}(O_k)})} \qquad (2)$$

where $C_i(O_k)$ is the count of codeword k in distribution i, $b_{i+j}(O_k)$ is the merged distribution by adding $b_i(O_k)$ and $b_j(O_k)$. Equation 2 measures the ratio between the probability that the individual distributions generated the training data and the probability that the merged distribution generated the training data in the similar manner as the generalized triphone.

Based on the similarity measure given in Equation 2, the Baum-Welch reestimation can be directly used to estimate the clustered distribution, which is consistent with the criterion used in our speaker-independent system. With speaker-dependent clustered distributions, the original speaker-independent models are interpolated. The interpolation weights can be either estimated using deleted interpolation or by mixing speaker-independent and speaker-dependent counts according to a pre-determined ratio that depends on the number of speaker-dependent data. Due to limited amount of adaptive data, the latter approach is more suitable to the former. It was also found that this procedure is more effective when the interpolation is performed directly on the raw data (counts), rather than on estimates of probability distributions derived from the counts. Before probability normalization, we have two sets of counts C_i^{s-dep} and $C_i^{s-indep}$ representing speaker-dependent and speaker-independent counts for distribution i. Let N_i denote the number of speaker-dependent data for distribution i. Final interpolated counts are computed with

$$C_i^{interpolated} = C_i^{s-indep} + log(1 + N_i) * C_i^{s-dep} \qquad (3)$$

from which we interpolate *interpolated* counts with context-independent models and uniform distributions with deleted interpolation. Varying the number of clustered distributions from 300 to 2100, speaker-adaptive recognition results are shown in Table 5. Just as in generalized triphone [12], the number of clustered distributions depends on the available adaptive data. From Table 5, we can see that when 40 sentences are used, the optimal number of clustered distributions is 500. The error rate is reduced from 3.6% (without distribution adaptation) to 3.1%. Detailed results for each speaker is shown in Table 6. In comparison with the speaker-independent system, the error reduction is more than 25%.

Speakers	Word Pair Grammar Error
BJW	2.4%
JLS	5.0%
JRM	4.5%
LPN	2.4%
Average	3.6%

Table 4: Detailed results using the SCHMM for each speaker.

Number of Clusters	Word-Pair Error Rate
300	3.2%
500	3.1%
900	3.3%
1500	3.3%
2100	3.4%

Table 5: Adaptation results with different clusters.

Speakers	Word Pair Error Rate
BJW	2.1%
JLS	4.6%
JRM	3.5%
LPN	2.4%
Average	3.1%

Table 6: Detailed results using 500 clusters for each speaker.

Incremental adaptation

The proposed algorithm can also be employed to incrementally adapt the voice of each speaker. Results are shown in Table 7. When 300 to 600 adaptive sentences are used, the error rate becomes lower than that of the best speaker-dependent systems. Here, we did not use clustered distributions because of available adaptation data. With 300-600 adaptive sentences, the error rate is reduced to 2.5-2.4%, which is better than the best speaker-dependent system trained with 600 sentences. This indicates speaker-adaptive speech recognition is quite robust since we could use information provided by speaker-independent models.

Conclusion and Future Work

In this paper, we used DARPA Resource Management task as our domain to investigate the performance of speaker-independent, speaker-

Incremental Sent	Word-Pair Error Rate
1	4.1%
40	3.6%
200	3.0%
300	2.5%
600	2.4%

Table 7: Incremental adaptation results.

dependent, and speaker-adaptive speech recognition. Not surprisingly, the error rate of speaker-dependent SPHINX is reduced by two to three times in comparison with the speaker-independent SPHINX, which demonstrated a great difference between speaker-dependent and -independent systems. Based on speaker-independent models, we studied speaker-adaptive speech recognition. With 40 adaptation sentences for each speaker, the error rate can be reduced from 4.3% to 3.1%. While the number of speaker adaptation sentences is comparable to that of speaker-dependent training, speaker-adaptive recognition works better than speaker-dependent recognition, which indicates the robustness of speaker-adaptive speech recognition. We also demonstrated that the same technique based on the maximum likelihood estimation criterion works well for speaker-independent, speaker-dependent, and speaker-adaptive speech recognition systems. For those different systems, emphasis should be focused on the available data and important free parameter sets.

In a different manner from model and codebook adaptation, we experimented with direct transformation of acoustic data. Direct normalization of cepstrum has achieved many successful results in environment adaptation [1]. Our normalization techniques involve cepstrum transformation of target speaker to the reference speaker. For each cepstrum vector X, two transformation matrix A and B are defined such that the SCHMM probability $Pr(AX + B|M)$ is maximized. The mapping structure used here can be regarded as a one-layer perceptron. The error rate for the same test set is only 3.9%, which is not a significant reduction in error rate. This indicates that the linear transformation used here is insufficient to bridge the difference between different speakers. Because of this, multi-layer perceptrons (MLP) with the back-propagation algorithm [19] are now employed for cepstrum transformation. A DTW algorithm is first used to warp the target data to the reference data. The input of the non-linear mapping network consists of vectors from the target speaker, and the optimal alignment pairs are used to supervise network learning. Initial results have shown that we can successfully minimize the difference between different speakers. We hope to further investigate this method.

Acknowledgements

This research was sponsored in part by US West and in part by the Defense Advanced Research Projects Agency (DOD), Arpa Order No. 5167, under contract number N00039-85-C-0163. The authors would like to express their gratitude to Raj Reddy, Richard Stern, and Hsiao-Wuen Hon at Carnegie Mellon for their comments and help.

References

[1] Acero, A. and Stern, R. *Environmental Robustness in Automatic Speech Recognition.* in: **ICASSP**. 1990, pp. 849–852.

[2] Brown, P. F., Lee, C.-H., and Spohr, J. C. *Bayesian Adaptation in Speech Recognition.* in: **ICASSP**. 1983, pp. 761–764.

[3] Feng, M., Kubala, F., and Schwartz, R. *Improved Speaker Adaptation Using Text Dependent Mappings.* in: **ICASSP**. 1988.

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

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

[6] Huang, X., Lee, K., and Hon, H. *On Semi-Continuous Hidden Markov Modeling.* in: **ICASSP**. 1990, pp. 689–692.

[7] Huang, X., Lee, K., Hon, H., and Hwang, M. *Improved Acoustic Modeling for the SPHINX Speech Recognition System.* in: **ICASSP**. 1991.

[8] Hwang, M., Hon, H., and Lee, K. *Between-Word Coarticulation Modeling for Continuous Speech Recognition.* Technical Report, Carnegie Mellon University, April 1989.

[9] Jelinek, F. and Mercer, R. *Interpolated Estimation of Markov Source Parameters from Sparse Data.* in: **Pattern Recognition in Practice**, edited by E. Gelsema and L. Kanal. North-Holland Publishing Company, Amsterdam, the Netherlands, 1980, pp. 381–397.

[10] Kubala, F. and Schwartz, R. *A New Paradigm for Speaker-Independent Training and Speaker Adaptation.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1990.

[11] Lee, C., Giachin, E., Rabiner, R., L. P., and Rosenberg, A. *Improved Acoustic Modeling for Continuous Speech Recognition.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1990.

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

[13] Lee, K. *Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition.* **IEEE Trans. on ASSP**, April 1990.

[14] Lee, K., Hon, H., and Reddy, R. *An Overview of the SPHINX Speech Recognition System.* **IEEE Trans. on ASSP**, January 1990, pp. 599–609.

[15] Nakamura, S. and Shikano, K. *Speaker Adaptation Applied to HMM and Neural Networks.* in: **ICASSP**. 1989.

[16] Nishimura, M. and Sugawara, K. *Speaker Adaptation Method for HMM-Based Speech Recognition.* in: **ICASSP**. 1988, pp. 207–211.

[17] Paul, D. *The Lincoln Robust Continuous Speech Recognizer.* in: **ICASSP**. 1989, pp. 449 – 452.

[18] Paul, D. and Martin, E. *Speaker Stress-Resistant Continuous Speech Recognition.* in: **ICASSP**. 1988.

[19] Rumelhart, D., Hinton, G., and Williams, R. *Learning Internal Representation by Error Propagation.* in: **Learning Internal Representation by Error Propagation**, by D. Rumelhart, G. Hinton, and R. Williams, edited by D. Rumelhart and J. McClelland. MIT Press, Cambridge, MA, 1986.

[20] Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., and Makhoul, J. *Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech.* in: **ICASSP**. 1985, pp. 1205–1208.

[21] Shikano, K., Lee, K., and Reddy, D. R. *Speaker Adaptation through Vector Quantization.* in: **ICASSP**. 1986.

[22] Stern, R. M. and Lasry, M. J. *Dynamic Speaker Adaptation for Isolated Letter Recognition Using MAP Estimation.* in: **ICASSP**. 1983, pp. 734–737.