A Study on Speaker-Adaptive Speech Recognition

X.D. Huang

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

Speaker-independent system is desirable in many applications where speaker-specific data do not exist. However, 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, DARPA Resource Management task is used as the domain to investigate the performance of speaker-adaptive speech recognition. Since adaptation is based on speaker-independent systems 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. Two parameter sets, the codebook mean vector and the output distribution, are regarded to be most important. They are modified in the framework of maximum likelihood estimation criterion according to the characteristics of each speaker. In order to reliably estimate those parameters, output distributions are shared with each other if they exhibit certain acoustic similarity. In addition to modify these parameters, speaker normalization with neural networks is also studied in the hope that acoustic data normalization will not only rapidly adapt the system but also enhance the robustness of speaker-independent speech recognition. Preliminary results indicate that speaker differences can be well minimized. In comparison with speaker-independent speech recognition, the error rate has been reduced from 4.3% to 3.1% by only using parameter adaptation techniques, with 40 adaptation sentences for each speaker. When the number of speaker adaptation sentences is comparable to that of speaker-dependent training, speaker-adaptive recognition works better than the best speaker-dependent recognition results on the same test set, which indicates the robustness of speaker-adaptive speech recognition.

1 INTRODUCTION

Speaker-independent speech recognition systems could provide users with a ready-to-use system [1, 2, 3, 4]. There is no 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. On the other hand, 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, half an hour of speech from the specific speaker is generally needed to reliably estimate 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 [2, 3]. A logical compromise for a practical system is to start with a speaker-independent system, and then adapt the system to each individual user.

Since adaptation is based on the speaker-independent system with only limited adaptation data, a good adaptation algorithm should be consistent with speaker-independent parameter estimation criterion, and adapt those parameters that are less sensitive to the limited training data. Two parameter sets, the codebook mean vector and the output distribution, are modified in the framework of maximum likelihood estimation criterion according to the characteristics of each speaker. In addition to modify those parameters, speaker normalization using neural networks is also studied in the hope that acoustic data normalization will not only rapidly adapt the system but also enhance the robustness of speaker-independent speech recognition.

The codebook mean vector can represent the essential characteristics of different speakers, and can be rapidly estimated with only limited training data [5, 6, 7]. Because of this, it is considered to be the most important parameter set. The semi-continuous hidden Markov model (SCHMM) [8] is a good tool to modify the codebook for each speaker. With robust speaker-independent models, the codebook is modified 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. Another important parameter set is the output distribution (weighting coefficients) of the SCHMM. Since there are too many parameters in the output distributions, direct use of the SCHMM would not lead to any improvement. The speaker-dependent output distributions are thus shared (by clustering) with each other if they exhibit certain acoustic similarity. Analogous to Bayesian learning [9], speaker-independent estimates can then be interpolated with the clustered speaker-dependent output distribution.

In addition to modify codebook and output distribution parameters, speaker normalization techniques are also studied in the hope that speaker normalization can not only adapt the system rapidly but also enhance the robustness of speaker-independent speech recognition [10]. Normalization of cepstrum has also achieved many successful results in environment adaptation [11]. The normalization techniques proposed here involve cepstrum transformation of any target speaker to

the reference speaker. For each cepstrum vector $\mathcal{X}$, the normalization function $\mathcal{F}(\mathcal{X})$ is defined such that the SCHMM probability $Pr(\mathcal{F}(\mathcal{X})|\mathcal{M})$ can be maximized, where $\mathcal{M}$ can be either speaker-independent, or speaker-dependent models; and $\mathcal{F}(\mathcal{X})$ can be either a simple function like $\mathcal{A}\mathcal{X}+\mathcal{B}$, or any complicated nonlinear function. Thus, a speaker-dependent function $\mathcal{F}(\mathcal{X})$ can be used to normalize the voice of any target speaker to a chosen reference speaker, or a speaker-independent function $\mathcal{F}(\mathcal{X})$ can be built to reduce speaker differences before speaker-independent training is involved such that the speaker-independent models are more accurate.

In this paper, DARPA Resource Management task is used as the domain to investigate the performance of speaker-adaptive speech recognition. An improved speaker-independent speech recognition system, SPHINX [12], is used as the baseline system here. The error rate for the RM2 test set, consisting of two male (JLS and LPN) and two female (BJW and JRM) speakers with 120 sentences for each, is 4.3%. This result is based on June 1990 system [13]. Recent results using the shared SCHMM is not included, which led to additional 15% error reduction [12].

Proposed techniques have been evaluated with the RM2 test set. With 40 adaptation sentences (randomly extracted from training set with triphone coverage around 20%) for each speaker, the parameter adaptation algorithms reduced the error rate to 3.1%. In comparison with the best speaker-independent result on the same test set, the error rate is reduced by more than 25% As the proposed algorithm can be used to incrementally adapt the speaker-independent system, the adaptation sentences is incrementally increased to 300-600. 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). For speaker normalization, two experiments were carried out. In the first experiment, two transformation matrix $\mathcal{A}$ and $\mathcal{B}$ are defined such that the speaker-independent SCHMM probability $Pr(\mathcal{A}\mathcal{X}+\mathcal{B}|\mathcal{M})$ is maximized. The error rate for the same test set with speaker-independent models is 3.9%. This indicates that the linear transformation is insufficient to bridge the difference among speakers. Because of this, the multi-layer perceptron (MLP) with the back-propagation algorithm [14, 15] is employed for cepstrum transformation. When the speaker-dependent model is used, the recognition error rate for other speakers is 41.9%, which indicates vast differences of different speakers. However, when 40 speaker-dependent training sentences are used to build the MLP, the error rate is reduced to 6.8%, which demonstrated the ability of MLP-based speaker normalization.

The paper is organized as follows. In Section 2, the baseline system for this study is described. Section 3 describes the techniques used for speaker-adaptive speech recognition, which consists of codebook adaptation, output distribution adaptation, and cepstrum normalization.

2 BASELINE SYSTEM

Large-vocabulary speaker-independent continuous speech recognition has made significant progress during the past years [1, 2, 3, 4]. Sphinx, a state-of-the-art speaker-independent speech recognition system developed at CMU [1], 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 [13]; (2) *generalized triphone models*. Triphones have been successfully used by [16, 17]. However, many contexts are quite similar, and can be combined. Clustering contexts leads to fewer, and thus more trainable, models [18]; (3) *function-word-dependent phone models*. These models were used to model phones in function words, which are typically short, poorly-articulated words such as *the, a, in, and*; (4) *between-word coarticulation modeling*. The concept of triphone modeling was extended to the word boundary, which leads to between-word triphone models [19]; (5) *semi-continuous models*. SCHMMs mutually optimize the VQ codebook and HMM parameters under a unified probabilistic framework [20], which greatly enhances the robustness in comparison with the discrete HMM [12]; (6) *speaker-clustered models*. Another advantage to use the SCHMM is that it requires less training data in comparison with the discrete HMM. Therefore, speaker-clustered models (male/female in this study) were employed to improve the recognition accuracy [12].

The above system was evaluated 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. Recent results using the shared distribution modeling have not yet included, which led to additional 15% error reduction [12].

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.

The same technology was extended for speaker-dependent speech recognition with 600/2400 training sentences for each speaker [21]. The SCHMM parameters and VQ codebook were estimated jointly starting with speaker-independent models. Results are listed in Table 2. The error rate of the speaker-dependent system can be reduced by three times in comparison with the speaker-independent system, albeit this comparison is not fair since the speaker-independent system is trained with 3990 sentences from about 100 speakers. However, these results clearly indicate the importance of speaker-dependent training data, and effects of speaker variability in the speaker-independent system. If speaker-dependent data or speaker-normalization techniques are available, the error rate may be significantly reduced.

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.

3 SPEAKER-ADAPTIVE SYSTEM

Last section clearly demonstrated the importance of speaker-dependent data, and requirements of speaker normalization mechanism for speaker-independent system design. This section will describe several techniques to adapt the speaker-independent system so that an initially speaker-independent system can be rapidly improved as a speaker uses the system. Speaker normalization techniques that may have a significant impact on both speaker-adaptive and speaker-independent speech recognition are also examined.

3.1 Codebook adaptation

The 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 [8, 20], In comparison with the conventional codebook adaptation techniques [5,6,7], 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 $Pr(\mathcal{X}|\mathcal{M})$, the probability of acoustic observations $\mathcal{X}$ given the model $\mathcal{M}$, is directly maximized. To elaborate, the posterior probability $\lambda_i(t)$ is first computed based on the speaker-independent model [20]. $\lambda_i(t)$ 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

$$\mu_i = \frac{\sum_t \lambda_i(t)\mathcal{X}_t}{\sum_t \lambda_i(t)} \quad (1)$$

In this study, the SCHMM is used 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. 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.

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.

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.

In fact, both the mean and variance vector can be adapted iteratively. However, the variances cannot be reliably estimated with limited adaptive data. Because of this, estimates are interpolated with speaker-independent estimates analogous to Bayesian adaptation [9, 22]. However, in comparison with iterative SCHMM codebook reestimation, there is no significant error reduction by combining interpolation into the codebook mapping procedure. It is sufficient by just using very few samples to reestimate the mean vector.

3.2 Output distribution adaptation

Several output-distribution adaptation techniques, including cooccurence mapping [23, 24], deleted interpolation [25, 20], and state-level-distribution clustering, are examined. 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 [24]. The output distribution of the speaker-independent models is then projected according to the cooccurence matrix. there is no 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 [23]. 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

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

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.

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.

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 is 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. Let C_i^{s-dep} and $C_i^{s-indep}$ represent speaker-dependent and speaker-independent counts for distribution i, $\mathcal{N}_i$ denote the number of speaker-dependent data for distribution i. Final interpolated counts are computed with

$$C_i^{interpolated} = C_i^{s-indep} + log(1 + \mathcal{N}_i) * C_i^{s-dep} \quad (3)$$

from which *interpolated* counts are interpolated 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 [23], the number of clustered distributions depends on the available adaptive data. From Table 5, it can be seen 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%.

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, clustered distributions are not used 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 information provided by speaker-independent models is available.

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.

3.3 Speaker normalization

Speaker normalization may have a significant impact on both speaker-adaptive and speaker-independent speech recognition. Normalization techniques proposed here involve cepstrum transformation of a target speaker to the reference speaker. For each cepstrum vector $\mathcal{X}$, the transformation function $\mathcal{F}(\mathcal{X})$ is defined such that the SCHMM probability $Pr(\mathcal{F}(\mathcal{X})|\mathcal{M})$ can be maximized, where $\mathcal{M}$ can be either speaker-independent or speaker-dependent models; and $\mathcal{F}(\mathcal{X})$ can be either a simple function as $\mathcal{A}\mathcal{X} + \mathcal{B}$ or any complicated nonlinear function. Thus, a speaker-dependent function $\mathcal{F}(\mathcal{X})$ can be used to normalize the voice of any target speaker to a chosen reference speaker for speaker-adaptive speech recognition. Furthermore, a speaker-independent function $\mathcal{F}(\mathcal{X})$ can also be built to reduce the difference of speakers before speaker-independent HMM training is applied such that the resulting speaker-independent models have sharp distributions.

In the first experiment, two transformation matrix $\mathcal{A}$ and $\mathcal{B}$ are defined such that the speaker-independent SCHMM probability $Pr(\mathcal{A}\mathcal{X} + \mathcal{B}|\mathcal{M})$ is maximized. The mapping structure used here can be regarded as a one-layer perceptron, where the SCHMM probability is used as the objective function. Based on the speaker-independent model, the error rate for the same test set is reduced from 4.3% to 3.9%. This indicates that the

linear transformation used here may be insufficient to bridge the difference between speakers.

As multi-layer perceptrons (MLP) can be used to approximate any nonlinear function, the fully-connected MLP as shown in Figure 1 is employed for speaker normalization. Such a network can be well trained with the back-propagation algorithm. The input of the nonlinear mapping network consists of three frames (3x13) from the target speaker. The output of the network is a normalized cepstrum frame, which is made to approximate the frame of the desired reference speaker. The objective function for network learning is to minimize the distortion (mean squared error) between the network output and the desired reference speaker frame. The network has two hidden layers, each of which has 20 hidden units. Each hidden unit is associated with a sigmoid function. For simplicity, the objective function used here has not been unified with the SCHMM. However, the extension should be straightforward.

To provide learning examples for the network, a DTW algorithm [26] is used to warp the target data to the reference data. Optimal alignment pairs are used to supervise network learning. For the given input frames, the desired output frame for network learning is the one paired by the middle input frame in DTW alignment. Since the goal here is to transform the target speaker to the reference speaker, the sigmoid function is not used for the output layer. Multiple input frames feeded to the network not only alleviate possible inaccuracy of DTW alignment but also incorporate dynamic information in the learning procedure. As nonlinear network may be less well trained, full connections between input units and output units are added. This has an effect of interpolation between the nonlinear network output and the original speech frames. This interpolation helps generalization capability of the nonlinear network significantly. To minimize the objective function, both nonlinear connection weights and direct linear connection weights are simultaneously adjusted with the back-propagation algorithm. Experimental experience indicates that 200 to 250 epochs are required to achieve acceptable distortion.

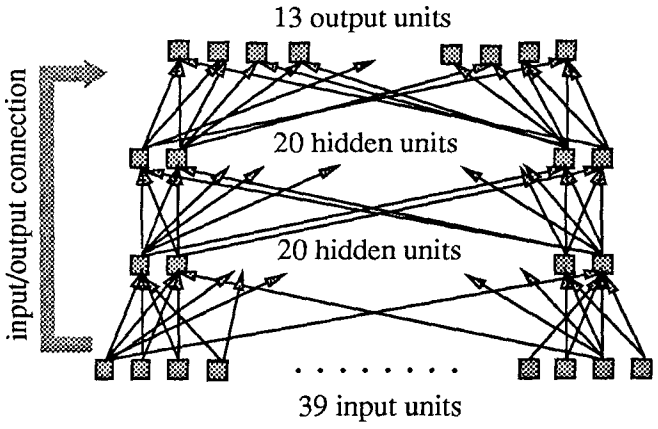

Figure 1: Speaker Net; 39 input units corresponding to 3 input frames, 13 output units corresponding to the normalized output frame

Since the study here is to investigate the capability of speaker normalization. Speaker-dependent models (2400 training sentences) are used instead of speaker-independent models. When the reference speaker is randomly selected as LPN, the average recognition error rate for the other three speakers is 41.9% as shown in Table 8. When 40 text-dependent training sentences are used to build the speaker normalization network, the average error rate is reduced to 6.8%. Note that neither codebook nor output distribution has been adapted yet in this experiment. The error rate has already been reduced by 80%. It is also interesting to note that for female speakers (JRM and BJW), speaker normalization dramatically reduces the error rate. Although the error rate of 6.8% is worse than that of the speaker-independent system (4.5%) for the same test set, this nevertheless demonstrated the ability of MLP-based speaker normalization.

Speakers	Word-Pair Error Without Normalization	Word-Pair Error With Normalization
JLS	8.5%	6.8%
BJW	62.1%	4.2%
JRM	55.3%	9.5%
Average	41.9%	6.8%

Table 8: Speaker normalization error rates.

4 DISCUSSION AND CONCLUSIONS

By using parameter adaptation techniques only, the error rate can be reduced from 4.3% to 3.1% with 40 adaptation sentences for each speaker. 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 the proposed speaker-adaptive speech recognition.

For speaker normalization, the error rate is reduced from 41.9% to 6.8% for cross speaker recognition with a speaker-dependent model. Here again, 40 training sentences are used

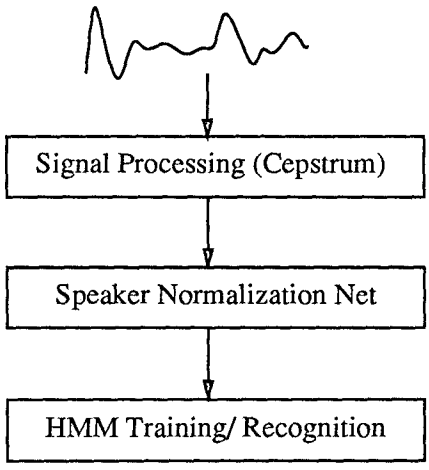

Figure 2: Speaker-independent speech recognition with speaker normalization network

to build the MLP-based nonlinear transformation function. The 80% error reduction demonstrated the ability of MLP-based speaker normalization. Due to the success of speaker normalization networks, a speaker-independent MLP-based network is being used as part of the front-end of the speaker-independent speech recognition system as shown in Figure 2. The network is built to reduce the difference of speakers before speaker-independent HMM training is involved such that speaker-independent models will have sharper distributions (better discrimination capability) in comparison with the conventional training procedure. Use of such normalization networks for speaker-independent speech recognition as well as unification of the SCHMM and MLP speaker normalization is currently in progress.

5 ACKNOWLEDGEMENTS

The author would like to express his gratitude to members of CMU speech group for their help; in particular, to Professor Raj Reddy and Dr. Kai-Fu Lee for their support.

References

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

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

[3] 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.

[4] 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.

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

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

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

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

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

[10] Kubala, F., Schwartz, R., and Barry, C. *Speaker Adaptation Using Multiple Reference Speakers.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1989.

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

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

[13] Huang, X., Alleva, F., Hayamizu, S., Hon, H., Hwang, M., and Lee, K. *Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition.* in: **DARPA Speech and Language Workshop**. Morgan Kaufmann Publishers, San Mateo, CA, 1990, pp. 327–331.

[14] 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.

[15] Lippmann, R. *Neural Nets for Computing.* in: **ICASSP**. 1988, pp. 1–6.

[16] 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.

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

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

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

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

[21] Huang, X. and Lee, K. *On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition.* in: **ICASSP**. 1991.

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

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

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

[25] 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.

[26] Sakoe, H. and Chiba, S. *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* **IEEE Trans. on ASSP**, vol. ASSP-26 (1978), pp. 43–49.