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

Xuedong Huang, *Member, IEEE*, and Kai-Fu Lee, *Member, IEEE*

Abstract—**Speaker-independent speech recognition systems are desirable in many applications where speaker-specific data do not exist. However, if speaker-dependent data become available, an originally speaker-independent system could be adapted to the specific speaker to further reduce the error rate. In this paper, we used the DARPA Resource Management task as our domain to investigate the performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. The error rate of our speaker-independent speech recognition system, SPHINX, was reduced substantially by incorporating between-word triphone models, additional dynamic features and sex-dependent semi-continuous hidden Markov models. The error rate for speaker-independent speech recognition is 4.3%. On speaker-dependent data, the error rate was further reduced to 2.6–1.4% with 600–2400 training sentences for each speaker, which demonstrated significant benefit of speaker-dependent training. Based on speaker-independent models, we studied speaker-adaptive speech recognition. Both codebooks and output distributions were considered for adaptation. We demonstrate that speaker-adaptive systems outperform both speaker-independent and speaker-dependent systems. This suggests that the most effective speech recognition system is the one that begins with speaker-independent training and continues to adapt to users.**

I. Introduction

MOST speech recognition systems are speaker-independent or speaker-dependent [12], [2], [18], [23], [15]. Speaker-independent systems require no training phase with data of users, and are desirable to many applications where training is difficult or impossible to conduct. However, because of between-speaker variabilities, well-trained speaker-dependent speech recognition systems outperform speaker-independent systems by a factor of two to three. Yet, to train a speaker-dependent system *well,* many hours of speech are usually needed. When the amount of speaker-dependent data is limited, such a performance improvement may not be realized. Therefore, one might argue that an ideal system is the one that begins with a speaker-independent system, and adapts to a speaker incrementally over time. One might also hope that such a system will outperform both speaker-independent and speaker-dependent systems, since it makes use of both speaker-independent and speaker-dependent databases. Therefore, a minimum amount of speaker-dependent training data may become sufficient. This papers describes such a system, and compares its performance to speaker-independent and speaker-dependent systems.

In this study, we used DARPA Resource Management (RM) task [24] as our domain to investigate the performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. The 997-word RM task is a database query task designed from 900 sentence templates. We used word-pair grammar that has a test-set perplexity of about 60. The training speech database consists of 3990 training sentences from 105 speakers for speaker-independent speech recognition; 600–2400 training sentences from each of four speakers (two males and two females) for speaker-dependent speech recognition (RM2, used in June 1990 evaluation [22]). The test set comprises a total of 480 sentences from the four speaker-dependent speakers. The testing speakers are not included in the speaker-independent training set. The speaker-independent training set includes 98% of the words in the vocabulary. The speaker-dependent training set includes 97% of the words in the vocabulary. When the number of speaker-dependent training sentences increased to 2400 for each speaker, 99% of the words are covered. The test set includes 73% of the words in the vocabulary. All these words appeared in the speaker-independent or speaker-dependent training set. For speaker-adaptive speech recognition, we used 40 sentences randomly extracted from the speaker-dependent training set. The word coverage is less than 19% for these 40 adaptation sentences. Both testing and training have the same recording conditions.

SPHINX was originally designed for continuous speaker-independent speech recognition [18]. Recently, the word recognition error rate was substantially reduced by incorporating between-word triphone models [11], high-order dynamic features [10], and sex-dependent semi-continuous hidden Markov models (SCHMM) [10]. For speaker-independent speech recognition, the error rate for the RM2 test set was reduced to 4.3%. Here, the error rate includes substitution, insertion, and deletion errors. This system gave the lowest error rate in June 1990 DARPA evaluation [22]. With the same technology, we extended SPHINX to speaker-dependent speech recognition, with 600 to 2400 speaker-dependent training sentences, the error rate was reduced to 2.6–1.4% (depending upon the amount of training data). Not surprisingly, the error rate is reduced by two to three times in comparison with the speaker-independent system.

Manuscript received January 17, 1991; revised April 13, 1992. This work was supported by the Defense Advanced Research Projects Agency (DOD) Arpa Order No. 5167 under Contract N00039–85-C-0163. The associate editor coordinating the review of this paper and approving it for publication was Prof. Mark A. Clements.

X. Huang is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.

K.-F. Lee is with Apple Computer, Inc., Cupertino, CA 951014.

IEEE Log Number 9206387.

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 vector quantization codebooks (or the SCHMM mixture components) and the output distributions (or the SCHMM mixing coefficients) in the framework of either discrete or semi-continuous models. Adaptation for codebooks and output distributions have been studied previously [26], [25], [4]. In these systems, there is no guarantee that incremental adaptation of both codebooks and output distributions will converge to a fully trained speaker-dependent system. We are interested in developing adaptation algorithms that are consistent with the estimation criterion used in either speaker-independent or speaker-dependent systems. In addition, the algorithm should adapt parameters that are less sensitive to limited training.

As demonstrated in speaker recognition experiments [27], the codebook can represent the essential characteristics of different speakers. Since the codebook mean vector can also be rapidly estimated with only a limited amount of training data, we consider it to be the most important parameter set. In SCHMM's [8], [5], we assume that the kth codeword is represented by a continuous Gaussian probability density function with mean $\boldsymbol{\mu}_k$ and diagonal covariance $\sum_k$. SCHMM's elegantly unify optimization of vector quantization codebooks and hidden Markov models (HMM) [8], which provides us with a good tool to modify the codebook for each speaker while holding the HMM parameters fixed. Based on speaker-independent models, we can modify the codebook so that the likelihood can be maximized for the given speaker. In the same manner as the speaker-independent or speaker-dependent system, this estimation procedure considers both phonetic and acoustic information in the estimation procedure. It gives us significant improvement even when the amount of adaptation data is limited.

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. Instead, we need to utilize information provided by the speaker-independent models. Analogous to Bayesian learning, we can interpolate speaker-dependent output distribution with speaker-independent estimates. Due to limited adaptive data, the similarity among output distributions of different phonetic models is measured so that different distributions can be grouped into clusters. Thus, interpolation is carried out between original speaker-independent distributions and clustered speaker-dependent distributions.

One desirable feature of our adaptation algorithm is that it converges well to speaker-dependent speech recognition. When our proposed algorithm is used to incrementally adapt the speaker-independent system, the number of optimal clustered distributions should be changed. As there is no easy way to determine the optimal number, we can simply interpolate output distribution without clustering output distributions. Although combination of codebook and output distribution adaptation gives us the best performance, it is interesting to note, as our empirical observation, that when the amount of adaptive data is limited, adapting codebook leads to substantial improvement; however, when the amount of available adaptive data is large, adapting output distributions becomes more important.

With 40 adaptation sentences (less than 19% word coverage) for each speaker, our adaptation algorithms reduced the error rate from 4.3% to 3.1%. This error reduction is more than 25% in comparison with the best speaker-independent system on the same test set. When we incrementally increased the adaptation sentences from 1 to 2400, the error rate was steadily reduced with the increase of the adaptation data. It was found that the error rate of speaker-adaptive speech recognition was always better or equal to that of speaker-dependent speech recognition trained with the same amount of data. With only 300 adaptation sentences, the error rate was the same as that of the speaker-dependent system trained with 600 sentences. This shows that speaker-adaptive speech recognition utilizes training data more effectively than speaker-dependent speech recognition.

This paper is organized as follows. In Section II, an improved speaker-independent speech recognition system is presented. In Section III, we extend our speaker-independent system for speaker-dependent speech recognition. Section IV discusses speaker-adaptive speech recognition. Our findings are summarized in Section V.

II. Speaker-Independent System

Significant progress has been made in large-vocabulary speaker-independent continuous speech recognition during the past years [18], [23], [14], [15]. Sphinx, a state-of-the-art speaker-independent speech recognition system developed at CMU [18], has achieved high word recognition accuracy. Recently, the error rate of the SPHINX system was further reduced by more than 50% with between-word coarticulation modeling [11], high-order dynamics [7], and sex-dependent SCHMM's [7]. This section will review the improved SPHINX system. The technology used here also forms the foundation for both speaker-dependent and speaker-adaptive speech recognition.

A. Signal Processing

The input speech signal is sampled at 16 kHz with a pre-emphasis filter, $1 - 0.95Z^{-1}$. Hamming window with a width of 20 ms is applied to speech signal every 10 ms. The 32-order LPC analysis is followed to compute the 12-order cepstral coefficients. Bilinear transformation of cepstral coefficients is employed to approximate a mel-scale representation. In addition, relative power energy is also computed together with cepstral coefficients. Speech features include the following.

1) LPC cepstral coefficients (dimension 12).
2) 40-ms and 80-ms differenced LPC cepstral coefficients (dimension 24)

$$\triangle \mathrm{cep}(t) = \mathrm{cep}(t+2) - \mathrm{cep}(t-2)$$
$$\triangle \mathrm{cep}'(t) = \mathrm{cep}(t+4) - \mathrm{cep}(t-4).$$

3) Second-order differenced cepstrum (dimension 12)

$$\triangle\triangle \mathrm{cep}(t) = \triangle \mathrm{cep}(t+1) - \triangle \mathrm{cep}(t-1).$$

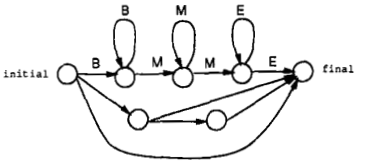


Fig. 1. The HMM topology used in SPHINX.

4) Power, 40-ms differenced power, second-order differenced power (dimension 3)

$$\Delta\text{power}(t) = \text{power}(t+2) - \text{power}(t-2)$$
$$\Delta\Delta\text{power}(t) = \Delta\text{power}(t+1) - \Delta\text{power}(t-1).$$

These features are vector quantized into four independent codebooks by the Linde-Buzo-Gray algorithm [19]. Each codebook has 256 entries.

B. Training Procedure

The phonetic HMM topology is shown in Fig. 1. There are three output distributions associated with the *arcs* for each HMM. They are labeled as Beginning, Middle, and Ending as illustrated in the figure. Unlabeled arcs are phoneme-dependent, where output distribution may be associated with different B, M or E.

Our training procedure is based on the forward–backward algorithm. Word models are formed by concatenating phonetic models; sentence models by concatenating word models. 48 context-independent *discrete* phonetic models are initially estimated from the uniform distribution. Deleted interpolation [13] is used to smooth estimated parameters with the uniform distribution. There are 7549 triphone models in the DARPA RM task when both within-word and between-word triphones are considered. Because of memory limitation, it is impossible to estimate all these triphone models. We started with the one-codebook system and estimated 7549 discrete models. The generalized-triphone clustering procedure [17] was then applied to reduce the number of models from 7549 to 1100. Here, the goal was to cluster similar triphones together such that these model parameters could be well trained.

Based on generalized triphone clusters, we first estimated 48 context-independent, four-codebook discrete models. With these context-independent models, we then estimated the 1100 generalized SCHMM's [7]. SCHMMs assume that the kth codeword is represented by a continuous probability density function $f_k(\boldsymbol{x})$, where $\boldsymbol{x}$ is the acoustic vector. The discrete output distribution $b_i(k)$ was replaced with the semi-continuous function $B_i(\boldsymbol{x})$:

$$B_i(\boldsymbol{x}) = \sum_{k=1}^{L} f_k(\boldsymbol{x}) b_i(k) \tag{1}$$

where L is the codebook size. In practice, 2 to 8 most significant $f_k(\boldsymbol{x})$'s are adequate. We also assume that each $f_k(\boldsymbol{x})$ is a Gaussian density function with mean $\boldsymbol{\mu}_k$ and diagonal covariance Σ_k. Means and covariance matrices were reestimated according to the following formula [8]:

$$\overline{\boldsymbol{\mu}_k} = \frac{\sum_t \sum_i \chi_t(i,k)\boldsymbol{x}_t}{\sum_t \sum_i \sum_k \chi_t(i,k)} \tag{2}$$

$$\overline{\Sigma_k} = \frac{\sum_t \sum_i \chi_t(i,k)(\boldsymbol{x}_t - \overline{\boldsymbol{\mu}_k})(\boldsymbol{x}_t - \overline{\boldsymbol{\mu}_k})^t}{\sum_t \sum_i \sum_k \chi_t(i,k)} \tag{3}$$

where $\chi_t(i,k)$ is the posterior probability that at time t, codeword k is emitted at Markov state i. For single codebook system, $\chi_t(i,k)$ can be computed as

$$\chi_t(i,k) = \sum_m \frac{\sum_j \alpha_{t-1}(i) a_{ij} b_j(O_k^t) f(\boldsymbol{x}_t \mid O_k^t) \beta_t(j)}{\Pr(\boldsymbol{X} \mid M)} \tag{4}$$

In (4), $\alpha()$ and $\beta()$ denote the forward and backward probabilities respectively; $b_j(\cdot)$ denotes output distributions associated with state j; a_{ij} denotes transition probability from state i to state j; $f(\boldsymbol{x}_t \mid O_k^t)$ denotes the Gaussian density for speech vector $\boldsymbol{x}_t$ at time t given the codeword O_k; m denotes the HMM index; and $\Pr(\boldsymbol{X} \mid M)$ denotes the probability of acoustic observation sequence, X, given the SCHMM, M. For a detailed treatment as well as multiple-codebook formula, readers are referred to [8].

Sex-dependent SCHMM's are used because of substantial differences between male and female speakers. In comparison with the discrete HMM, the smoothing capability of the SCHMM alleviates the problem of data fragmentation.

C. Recognition Procedure

For each input utterance, the Viterbi beam search algorithm was used to find out the optimal state sequence in the language network. In order to use sex-dependent SCHMM's, codebook-based sex classification [6], [27] was carried out before recognition started. Experiments show that the error rate of sex-classification is below 1%. Based on sex-classification, only the corresponding sex-dependent SCHMM's were activated for the Viterbi search.

D. Speaker-Independent Results

We evaluated our new system on the June 1990 (RM2) test set using the word-pair grammar, as shown in Fig. 2. When between-word triphone models were used, the error rate was reduced by 20% over the original SPHINX. When additional dynamic features were incorporated in the multi-codebook framework, the error rate was reduced by another 15%. Finally, the sex-dependent SCHMM further reduced the error rate by 20%. When these techniques were combined, overall error rate reduction was more than 50% in comparison with the original Sphinx system [18]. When no grammar is used, similar improvements were observed [10]. Thus in this study, we will evaluate our systems using the word-pair grammar only.

Based on the improved system, the word recognition error rates of RM2 for each speaker are shown in Table I. The

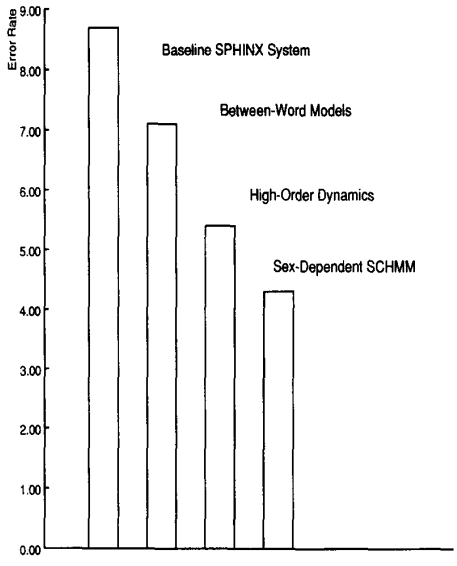


Fig. 2. Improvements for speaker-independent speech recognition.

TABLE I
SPEAKER-INDEPENDENT RESULTS

Speaker	3990 Training Sentences Word Error Rate
BJW	3.1%
JLS	4.8%
JRM	5.8%
LPN	3.6%
Average	4.3%

average error rate is 4.3%, which gave the lowest error rate in June 1990 DARPA evaluation [22]. The improved system will be referred to as the baseline system in comparison with both speaker-dependent and speaker-adaptive speech recognition.

III. SPEAKER-DEPENDENT SYSTEM

It is generally agreed that a speaker-dependent system outperforms a speaker-independent system if a sufficient amount of training data is available to obtain speaker-dependent parameters. In this section, we will extend the SPHINX system for speaker-dependent speech recognition. We will see that technology developed for speaker-independent speech recognition can be well applied to speaker-dependent speech recognition.

A. System Overview

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. We used generalized triphone models derived from speaker-independent training set. Based on speaker-dependent data, the SCHMM parameters and VQ codebook are estimated jointly starting with sex-dependent models and codebooks. Two to three iterations were run to sharpen the output distributions. Although sharpened distributions may lead to degraded performance for speaker-independent speech recognition, we observed that speaker-dependent speech recognition generally results in better performance.

TABLE II
SPEAKER-DEPENDENT RESULTS

Speaker	600 Training Sentences Word Error Rate	2400 Training Sentences Word 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%

B. Speaker-Dependent Results

Results are listed in Table II. 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 this speaker-dependent system. One should be careful about this comparison, as both systems can be improved by adding more training data. However, these results clearly indicate the importance of speaker-dependent 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.

IV. SPEAKER-ADAPTIVE SYSTEM

While last section clearly demonstrated the importance of speaker-dependent data, it is generally impractical for a speaker to speak 2400 sentences (more than 2 h) 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. We are also interested in developing adaptation algorithms that at least converge to the fully trained speaker-dependent system and can incrementally improve speaker-independent performance as a speaker uses the system. We will examine how to adapt two most important parameter sets: the codebook (or the SCHMM mixture components) and the output distributions (or the SCHMM mixing coefficients).

A. Codebook Adaptation

As mentioned in Section II, the SCHMM has been used to extend the discrete HMM by replacing discrete output distributions with a combination of the original discrete output distributions and continuous density function of a codeword [9], [8]. In comparison with the conventional codebook adaptation techniques [26], [21], [20], the SCHMM codebook and SCHMM parameters can be jointly optimized according

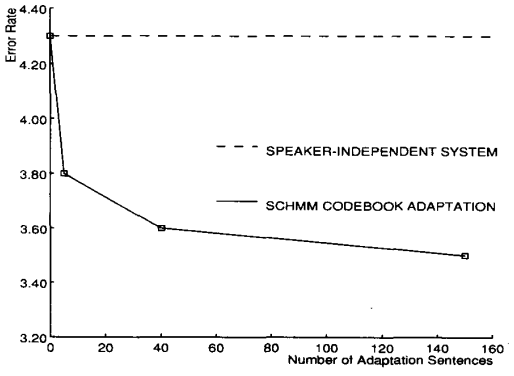


Fig. 3. Codebook adaptation results using the SCHMM.

to the maximum-likelihood criterion. Thus the SCHMM can be readily applied to speaker-adaptive speech recognition by reestimating the codebook parameter while holding the HMM parameters fixed.

In this study, we used speaker-independent models as initial models to modify the codebook 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 $\Pr(\mathrm{X} \mid \mathrm{M})$ is directly maximized. Here, we assume that each codeword, $f_k(\boldsymbol{x})$, is a Gaussian density function with mean $\boldsymbol{\mu}_k$ and diagonal covariance $\sum_k$. Both the mean vector and variance matrices can be adapted iteratively with (2) and (3). However, the variances cannot be reliably estimated with a limited amount of adaptive data. Because of this, we tried to interpolate estimates with speaker-independent estimates analogous to Bayesian adaptation [1], [28]. Unfortunately, in comparison with iterative SCHMM codebook reestimation, we have not achieved any significant error reduction. Using very few samples seems sufficient to reestimate the mean vector.

Varying the number of adaptation sentences from 5 to 150 for each speaker, speaker-adaptive recognition results are shown in Fig. 3. Detailed results using 40 adaptive sentences are also listed in Table III. In these experiments, we used the SCHMM to reestimate the mean vector only. Three iterations were carried out for each speaker. The error rates with 5 to 40 adaptive sentences are 3.8% and 3.6%, respectively. In comparison with the speaker-independent model, the error rate of adaptive systems is reduced by more than 15% with only 40 sentences from each speaker. Further increase in the number of adaptive sentences did not lead to any significant improvement in spite of the substantial word coverage increase.

B. Output Distribution Adaptation

Another set of parameters that could be adapted is the output distributions. The present problem with output distribution adaptation is the large number of parameters. Several techniques that tried to combat the parameter estimation problems include cooccurence mapping [16], [3], deleted interpolation [13], [8], and state-level-distribution clustering. All these studies are based on the SCHMM-adapted codebook as discussed above.

TABLE III
DETAILED CODEBOOK ADAPTATION RESULTS USING 40 SENTENCES FROM EACH SPEAKER

Speaker	Word Error Rate
BJW	2.4%
JLS	5.0%
JRM	4.5%
LPN	2.4%
Average	3.6%

In cooccurence mapping, the cooccurence matrix provides the probability of codewords of the target speaker given the codeword of speaker-independent models [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 mapping only plays a role of smoothing, which is particularly suitable to adapt from speaker-dependent models [3].

A better adaptation technique should be consistent with the criterion used in the original speech recognition system. Therefore, the approach based on maximum-likelihood estimation should be used. In our current system, there are total 3300 output distributions. 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 estimation problem, we take advantage of the similarity between output distributions of different phonetic models. If two distributions are similar, they are grouped into the same cluster. Since clustering is carried out at the state-level, it is more flexible and more reliable in comparison with model-level clustering. This is because clustering two entire models may force output distributions with quite different shapes to be merged together when *only parts* of the models exhibit close resemblance.

To cluster two output 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

$$d(b_i, b_j) = \frac{\left(\prod_k b_i(O_k)^{C_i(O_k)}\right)\left(\prod_k b_j(O_k)^{C_j(O_k)}\right)}{\left(\prod_k b_{i+j}(O_k)^{C_{i+j}(O_k)}\right)} \quad (5)$$

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 $b_j(O_k)$ (with normalization). Equation (5) 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. With such a similarity measure, the clustering algorithm can be used as follows:

1) All HMM's are first estimated.
2) Initially, every distribution of all HMM's is created as a cluster.
3) Find the most similar pair of clusters and merge them together.
4) For each pair of clusters, consider moving every element from one to the other:

a) move the element if the resulting configuration is an improvement;
b) repeat until no such moves are left.

5) Go to step 3 unless some convergence criterion is met.

We first derived clustered distributions based on speaker-independent models. Here, clustering can be carried out for all the distributions without any constraints. We could also only allow those distributions from the same context-independent phonetic models to be merged. Given the desired total number of distributions, the number of distributions from each context-independent phonetic models was determined by the overall distortion. It was found that both approaches produced similar results, but the latter was computationally more efficient. Once clustering was done, these clustered distributions were fixed for subsequent speaker-dependent training.

Armed with distribution-clustering, the Baum—Welch reestimation were used to estimate the clustered distribution, which is consistent with the criterion used in our speaker-independent system. To apply the forward—backward algorithm, parameter counts for clustered distributions need to be accumulated before Baum—Welch reestimation. In the same manner as the SCHMM, parameter-sharing will not affect the maximum-likelihood estimation criterion [8]. The Q-function can be modified to prove this [8].

It is easy to map clustered distributions back to original distributions. Thus the original speaker-independent distributions can be interpolated with clustered speaker-dependent distributions. 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 amount of speaker-dependent data. Due to limited adaptation data, the latter approach is more suitable. It was also found that this procedure was more effective when the interpolation was 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^{sd} and C_i^{si} representing speaker-dependent and speaker-independent counts for distribution i. Let N_i denote the number of speaker-dependent tokens observed for distribution i. Final interpolated counts were computed with

$$C_i^{\text{interpolated}} = C_i^{\text{si}} + \log(1 + N_i) * C_i^{\text{sd}} \tag{6}$$

from which we then interpolated $C_i^{\text{interpolated}}$ with corresponding context-independent distributions and uniform distributions using deleted interpolation [13]. Here, the factor $\log(1 + N_i)$ was determined experimentally.

Varying the number of clustered distributions from 300 to 2100, speaker-adaptive recognition results are shown in Fig. 4. In a manner similar to generalized triphone [16], the number of clustered distributions depends on the amount of available adaptive data. When the amount of adaptive data changes, the optimal number will also change, although the difference may not be significant. As shown in Fig. 4, when 40 sentences are used, we can see that the best performance is obtained when the number of clusters is around 500. Here, the total number of original distributions is 3300.

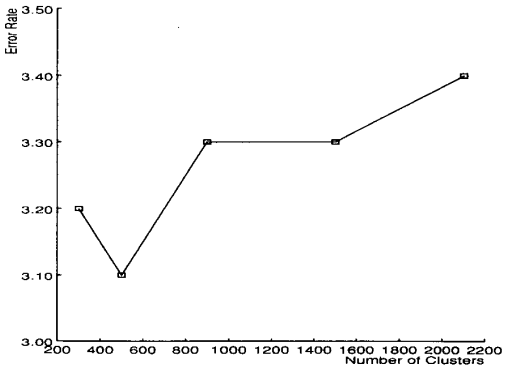


Fig. 4. Output distribution adaptation results using different distribution clusters.

TABLE IV
OUTPUT DISTRIBUTION ADAPTATION RESULTS USING 500 DISTRIBUTION CLUSTERS FOR EACH SPEAKER

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

Using 500 clusters, detailed results for each speaker are listed in Table IV. The average error rate is reduced from 3.6% (without distribution adaptation) to 3.1%. In comparison with the speaker-independent speech recognition system, the error reduction is more than 25%. We can also see that the error rates are consistently reduced across four speakers.

C. Incremental Adaptation

The proposed algorithm can be employed to incrementally adapt the voice of each speaker. This is feasible in real applications during the usage of the speech recognition system. Here, we need to know the right text before carrying out adaptation. We did not use clustered distributions since the number of clustered distributions must be determined according to the amount of available adaptation data. Instead, (6) was used to interpolate un-clustered speaker-dependent distributions with original speaker-independent distributions. Deleted interpolation was then used to interpolate context-dependent and context-independent models. The SCHMM codebook was also adapted as described above. One attractive features of our adaptation algorithm is that it converges well to speaker-dependent speech recognition. SCHMM codebook optimization is of course consistent with the estimation criterion used in both speaker-dependent and speaker-independent systems. For output distribution interpolation, the first term in (6) will be negligible with the increase of N_i as the contribution of the second term will dominate $C_i^{\text{interpolated}}$.

When we gradually increased the number of adaptation sentences, the error rate was steadily reduced. Incremental adaptation results are shown in Table V. When the number

TABLE V
INCREMENTAL ADAPTATION RESULTS

Incremental Sentences	Word Error Rate
0	4.3%
1	4.1%
40	3.6%
200	3.0%
300	2.5%
600	2.4%
2400	1.4%

of training sentences is less than 600, the error rate was lower than that of the best speaker-dependent systems trained with the same amount of data. When the number of adaptive sentences reached 2400, speaker-adaptive speech recognition converged to speaker-dependent speech recognition. It is also interesting to note that when the amount of adaptive data is limited, adapting codebook leads to substantial improvement; however, when the amount of available adaptive data is large, adapting output distributions becomes more important. In fact, when 2400 adaptive sentences were used, the performance degradation with fixed SCHMM codebook (estimated using 40 sentences only) was less than 5%.

As shown in Fig. 5, when compared with fully trained speaker-dependent speech recognition, we can see that the error rate with speaker-adaptive speech recognition is always equal to or lower than that with speaker-dependent speech recognition. When the amount of speaker-dependent training data is limited, the performance of speaker-dependent speech recognition is significantly worse than that of speaker-independent speech recognition. However, the performance of speaker-dependent speech recognition can be substantially improved with adaptation data. Likewise, speaker-adaptive speech recognition benefits from more speaker-dependent data, but they can also make use of the information provided by speaker-independent models. In all cases, the error rate of speaker-adaptive speech recognition is always equal to or less than that of either speaker-dependent or speaker-independent speech recognition.

V. SUMMARY

In this paper, we used DARPA Resource Management task as our domain to investigate the performance of large-vocabulary continuous speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. We 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. We believe that an ideal system is the one that begin, with a speaker-independent system, and adapts to a speaker incrementally over time. We demonstrated that such a system outperform both speaker-independent and speaker-dependent systems, since it makes use of information existing in both speaker-independent and speaker-dependent databases.

For speaker-independent speech recognition, the error rate of the original SPHINX system was significantly reduced by incorporating high-order dynamic features and sex-dependent

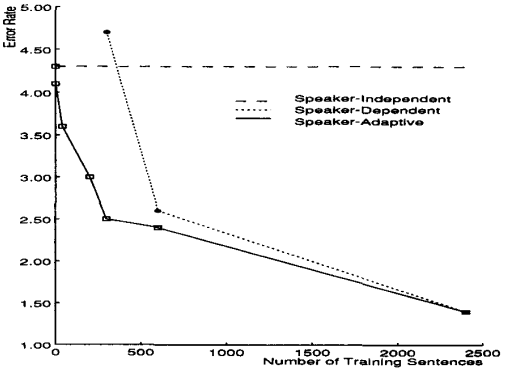


Fig. 5. Performance of speaker-independent, speaker-dependent, and speaker-adaptive speech recognition.

SCHMMs. For speaker-dependent speech recognition, the error rate was further reduced by additional two to three times in comparison with the speaker-independent system. This demonstrated the necessity for speaker-adaptive speech recognition. Based on speaker-independent models, we studied speaker-adaptive speech recognition using SCHMM codebook mapping and output distribution interpolation. With 40 adaptation sentences for each speaker, the error rate was reduced from 4.3% to 3.1%. When the amount of adaptation data is limited, adapting codebook can lead to substantial improvement. However, when the amount of available adaptation data is large, adapting output distributions becomes more important. Speaker-adaptive recognition always performs as well as or better than both speaker-dependent recognition and speaker-independent speech recognition.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Raj Reddy for his support and encouragement and to H.W. Hon and M.Y. Hwang for their help and comments.

REFERENCES

[1] P. F. Brown, C.-H. Lee, and J. C. Spohr, "Bayesian adaptation in speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 761–764, 1983.

[2] Y. L. Chow, R. Schwartz, S. Roucos, O. Kimball, P. Price, F. Kubala, M. Dunham, M. Krasner, and J. Makhoul, "The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1986.

[3] M. Feng, F. Kubala, and R. Schwartz, "Improved speaker adaptation using text dependent mappings," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 131–134, 1988.

[4] M. Ferretti and S. Scarci, "Large-vocabulary speech recognition with speaker-adapted codebook and HMM parameters," in *Proc. Eurospeech*, pp. 154–156, 1989.

[5] X. Huang, "Phoneme classification using semi-continuous hidden Markov models," *IEEE Trans. Signal Processing*, vol., pp. 1062–1067, May 1992.

[6] ——, "A study on speaker-adaptive speech recognition," in *DARPA Speech and Language Workshop*. San Mateo, CA, 1991.

[7] X. Huang, F. Alleva, S. Hayamizu, H. Hon, M. Hwang, and K. Lee, "Improved hidden Markov modeling for speaker-independent continuous speech recognition," in *DARPA Speech and Language Workshop*, pp. 327–331, 1990.

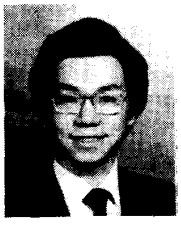
[8] X. Huang, Y. Ariki, and M. Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh, U.K.: Edinburgh Univ. Press, 1990.

[9] X. Huang and M. Jack, "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, vol. 3, pp. 239–252, 1989.
[10] X. Huang, K. Lee, H. Hon, and M. Hwang, "Improved acoustic modeling for the SPHINX speech recognition system," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toronto, Ont., Canada, pp. 345–348, 1991.
[11] M. Hwang, H. Hon, and K. Lee, "Modeling between-word coarticulation in continuous speech recognition," in *Proc. Eurospeech*, Paris, France, pp. 5–8, 1989.
[12] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, vol. 73, pp. 1616–1624, 1985.
[13] F. Jelinek and R. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Pattern Recognition in Practice*, E. Gelsema and L. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, pp. 381–397, 1980.
[14] F. Kubala and R. Schwartz, "A new paradigm for speaker-independent training and speaker adaptation," in *DARPA Speech and Language Workshop*, 1990.
[15] C. Lee, E. Giachin, R. Rabiner, and A. Rosenberg, "Improved acoustic modeling for continuous speech recognition," in *DARPA Speech and Language Workshop*, 1990.
[16] K. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*. Boston, MA: Kluwer Academic, 1989.
[17] —, "Context-dependent phonetic hidden Markov models for continuous speech recognition," *IEEE Trans. Signal Processing*, pp. 599–609, vol. Apr. 1990.
[18] K. Lee, H. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Trans. Signal Processing*,vol., pp. 35–45, Jan. 1990.
[19] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.
[20] S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 89–92, 1989.
[21] M. Nishimura and K. Sugawara, "Speaker adaptation method for HMM-based speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 207–211, 1988.
[22] D. Pallett, J. Fiscus, and J. Garofolo, "DARPA resource management benchmark test results June 1990," in *DARPA Speech and Language Workshop*, pp. 298–305, 1990.
[23] D. Paul, "The Lincoln robust continuous speech recognizer," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 449–452, 1989.
[24] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "A database for continuous speech recognition in a 1000-word domain," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 651–654, 1988.
[25] R. Schwartz, Y. Chow, and F. Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1987.
[26] K. Shikano, K. Lee, and D. R. Reddy, "Speaker adaptation through vector quantization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 2643–2646, 1986.
[27] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 387–390, 1985.
[28] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for isolated letter recognition using MAP estimation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 734–737, 1983.

Xuedong Huang (M'89) was born in China on October 20, 1962. He received the B.Sc. degree in Computer Science from Hunan University, Changsha, China in 1982, the M.Sc. degree in Computer Science from Tsinghua University, Beijing, China in 1984, and the Ph.D. degree in Electrical Engineering from the University of Edinburgh, Edinburgh, U.K. in 1989.

From 1982 through 1986, he worked on Chinese voice dictation at Tsinghua University, and received the Science and Technology Progress Awards from National Education Commission of China in 1987. During 1987–1989, he was with the University of Edinburgh, where he worked on adaptive signal processing and originated semi-continuous hidden Markov models. In 1989, he joined the faculty of Carnegie Mellon University, where he is currently a Research Computer Scientist and is directing speech recognition effort within the spoken language understanding group at Carnegie Mellon. His current interests lie in speech recognition, spoken language systems, stochastic language modeling, neural networks, and digital signal processing. He has published one book (*Hidden Markov Models for Speech Recognition*, Edinburgh, Scotland: Edinburgh Univ. Press, 1990), as well as over 40 articles in speech recognition, neural networks, and signal processing.

Dr. Huang is a member of Sigma Xi.



Kai-Fu Lee (S'85–M'88) was born in Taipei, Taiwan in 1961. He received the A.B. degree (summa cum laude) in Computer Science from Columbia University, New York, in 1983, and the Ph.D. degree in Computer Science from Carnegie Mellon University, Pittsburgh, PA, in 1988.

From 1988 to 1990, Dr. Lee directed the speech recognition effort within the spoken language understanding group at Carnegie Mellon, where he was a Research Computer Scientist and Assistant Professor. Since 1990, he has been a Principal Scientist at Apple Computer, Inc. Currently, he manages the Apple's Speech and Language Technologies Group. His current research interests include automatic speech recognition, stochastic modeling, pattern recognition, speech synthesis, natural language processing, spoken language systems, and neural networks. He has published two books: *Automatic Speech Recognition* (Boston, MA: Kluwer Academic, 1989), and *Readings in Speech Recognition* (Morgan Kaufman 1990 co-edited with Alex Waibel).

Dr. Lee is a member of Phi Beta Kappa, Sigma Xi, and the Acoustical Society of America. He received the Best Paper Award from the IEEE Signal Processing Society in 1991.