

# Selective Training for Hidden Markov Models with Applications to Speech Classification

Levent M. Arslan, *Member, IEEE*, and John H. L. Hansen, *Senior Member, IEEE*

**Abstract**—Traditional maximum likelihood estimation of hidden Markov model parameters aims at maximizing the overall probability across the training tokens of a given speech unit. As such, it disregards any interaction or biases across the models in the training procedure. Often, the resulting model parameters do not result in minimum error classification in the training set. A new selective training method is proposed that controls the influence of outliers in the training data on the generated models. The resulting models are shown to possess feature statistics which are more clearly separated for confusable patterns. The proposed selective training procedure is used for hidden Markov model training, with application to foreign accent classification, language identification, and speech recognition using the *E*-set alphabet. The resulting error rates are measurably improved over traditional forward-backward training under open test conditions. The proposed method is similar in terms of its goal to maximum mutual information estimation training, however it requires less computation, and the convergence properties of maximum likelihood estimation are retained in the new formulation.

**Index Terms**—Classifier training, hidden Markov models, pattern classification, speech recognition.

## I. INTRODUCTION

FOR SPEECH recognition, the two most popular frameworks that require rigorous training strategies are hidden Markov models (HMMs) [23], and artificial neural networks (ANNs) [17]. In general, HMMs are preferred over neural networks, because their implementations are simpler, faster, and they generally require less training data than ANNs for most recognition tasks. Recently, HMM-ANN hybrids have been proposed which combine both modeling strategies in order to improve performance [5], [18].

The performance of a recognition system depends heavily on the complexity of the vocabulary. For example, the ability of a recognizer to distinguish between “white” and “wide” is much more limited than distinguishing between “hot” and “destination.” Such issues in vocabulary confusability and

complexity have motivated studies that attempt to increase the separability among similar speech patterns. Linear discriminant analysis [7] is one method of transforming and scaling variables to improve the performance of a classification system. It was first successfully applied to speech recognition by Hunt [13] in independent mel-scale linear discriminant analysis (IMELDA). Various studies have shown improvement in speech recognition using this technique with subword models [3], [9]. Further refinements of this technique were later developed by Ayer [6] for use in whole-word recognition, and also by Parris [22] to incorporate state specific mixture densities. Other discriminative training methods have also been proposed in order to correct for classification errors in the training set [4], [8], [20], [21]. Bahl *et al.* [4] proposed the maximum mutual information estimation (MMIE) technique, which increases the *a posteriori* probability of the model corresponding to the training data, given the data. Gopalakrishnan *et al.* [8] later introduced a reestimation formula for discrete HMMs that applies to rational objective functions. Normandin *et al.* [20], [21] showed significant improvements in a connected digit recognition task using the MMIE technique with an extension of the formulation to include the continuous density case. Although Merialdo [16] suggested some improvement to the convergence properties of the MMIE technique, unlike maximum likelihood estimation, there is no known reestimation formulation for MMIE training with theoretically proven convergence. Recently, Juang and Katagiri [14] proposed another technique that minimizes the number of errors in the training set by weighing the feature set. They demonstrated marked improvement using this method for the highly confusable *E*-set. The focus of our approach is similar in spirit to the approaches mentioned, since the goal is to improve overall recognition performance given the available training data. However, discriminative training methods generally focus on increasing the separable distance between models, normally their means. As a result, they focus on changes to the resulting model itself. The method proposed here, termed *selective training*, makes a departure from previous methods since it does not always force the models to fit the training data, but rather deemphasizes that data which does not fit the models well. It is worth mentioning at this point that if the training data is collected and labeled perfectly, the proposed selective training method can also be used to emphasize error tokens in the training data. In other words, to place more training weight on the outliers, since this will improve performance for tasks such as language ID and the confusable *E*-set task (because the data is labeled

Manuscript received August 31, 1996; revised February 26, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

L. M. Arslan was with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA. He is now with the Electrical and Electronics Engineering Department, Boğazici University, Istanbul, Turkey.

J. H. L. Hansen was with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA. He is now with the Robust Speech Processing Laboratory and the Center for Spoken Language Understanding, University of Colorado, Boulder, CO 80309 USA (e-mail: jhlh@ee.duke.edu; <http://www.ee.duke.edu/Research/Speech>).

Publisher Item Identifier S 1063-6676(99)00172-8.

perfectly). The proposed technique is also simple to implement and quite fast. Only a few added iterations to that needed in the traditional forward-backward algorithm are necessary.

The outline of this paper is as follows. In Section II, the proposed selective training method is presented with a complete discussion of the model update equations. Section III presents an extensive set of evaluations using selective training. Evaluations are illustrated for simulated and real data. Finally, Section IV presents summary and draws conclusions.

## II. SELECTIVE TRAINING

When developing effective speech recognition systems, it is typically necessary to use large amounts of training data. This is especially true for applications requiring speaker independence or large vocabularies. When the training data corpus is large, it is often difficult to guarantee that automatic labeling is done properly for the entire data set. As a result, a single label error can cause the resulting HMM to have inappropriate model means or covariances. As an example, a segment that is labeled as the /s/ phoneme may in fact correspond to an /f/ phoneme (for example “six” versus “fix”). Naturally, if we assume that all the tokens for this training set here belong to the same phoneme, then the incorrectly labeled token will force a significant shift in the HMM model structure. The problem here is that as speech recognition applications become more advanced, larger amounts of speech data are used in the training phase, making it more difficult to reliably verify that the training data is correctly labeled.

The impact of incorrectly labeled training tokens became very apparent in a recent study by Arslan and Hansen, which considered the problem of foreign accent classification [10]. Since the problem was focused on accent assessment, it was assumed that all speakers who indicated their native language to be non-English should always be classified as that particular accent. In reality, speakers vary in the degree of accent they display depending on their experience in the second language. As such, it was necessary to consider a procedure that would reduce the impact of speakers who did not display strong accent traits when training a particular accent model. In the training phase, words from speakers who have a common native language were employed to form accented word models for that language. However, depending on various factors (e.g., the second language learning age, length of residence in second language speaking country, etc.), nonnative speakers included in the data base had varying degrees of accent, while some speakers were able to utter some words from the selected vocabulary with neutral American English pronunciation patterns. Under these circumstances, it is better to suppress the weight of those word tokens that are not representative of the language accent, or if possible, not to use them at all in training language accent models. If these misclassified tokens were employed directly in a standard training method, the result would be a biased shift in the model mean and covariance matrix for the accented word model.

In order to explain the motivation behind the proposed selective training method, a hypothetical classification problem is considered. In Fig. 1, an example scatter plot of two-

## SELECTIVE TRAINING ILLUSTRATION

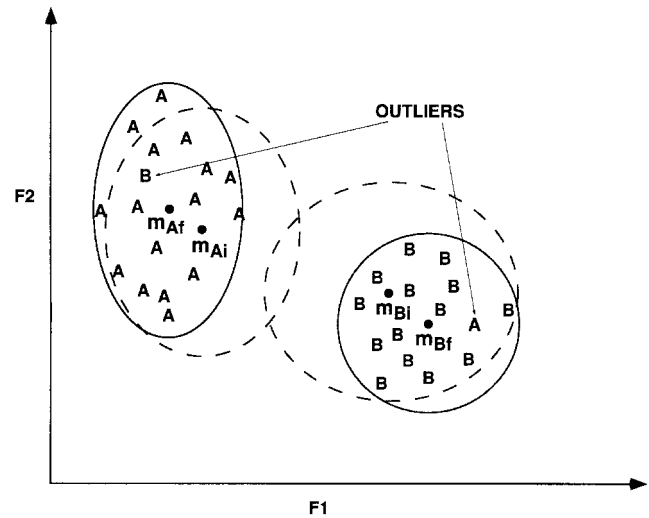


Fig. 1. Illustration of the adverse influence of outliers when creating statistical models for different classes.

dimensional (2-D) feature vectors is shown. The aim here is to distinguish between classes labeled as A and B. Therefore, statistical models for the two classes must first be generated based on the training data that is shown in the figure. The usual recognition approach is to represent the two classes with 2-D Gaussian densities with means and variances computed from corresponding class samples. Following this approach, the initial means from input training data for classes A and B are found to be  $m_{Ai}$  and  $m_{Bi}$ , respectively. The variances are represented by dashed ellipses for the same classes. However, as illustrated in this figure, an outlier exists for both classes. These outliers would result in errors if tested with the models that were just generated. Naturally, if these outliers were originally labeled correctly, it would be better to use the existing models. However if the labeling procedure is prone to errors, it may be better to exclude these outliers, or reduce their influence in the training process to estimate more accurate models. In this example, when the outliers are excluded, the means for the two classes shift to  $m_{Af}$  and  $m_{Bf}$ , with corresponding new variances represented as the solid ellipses. It is clear from this example that the new models can better characterize the overall statistics when these outliers are excluded.

This approach can be extended to the training of HMM's from labeled data. This is especially important for such applications as foreign accent classification, where the labeling of speech data may be unreliable depending on the level of accent exhibited by each speaker. One approach would be as follows. First, generate models from the given training data assuming it is correctly labeled. Next, using these models identify outliers by testing the available training data and determine where significant errors occur. Finally, retrain the models using the same training data by either adjusting the weight of the outliers or eliminating them in the training process.

Another approach is to weight the training tokens according to their relative match for their own word models and their

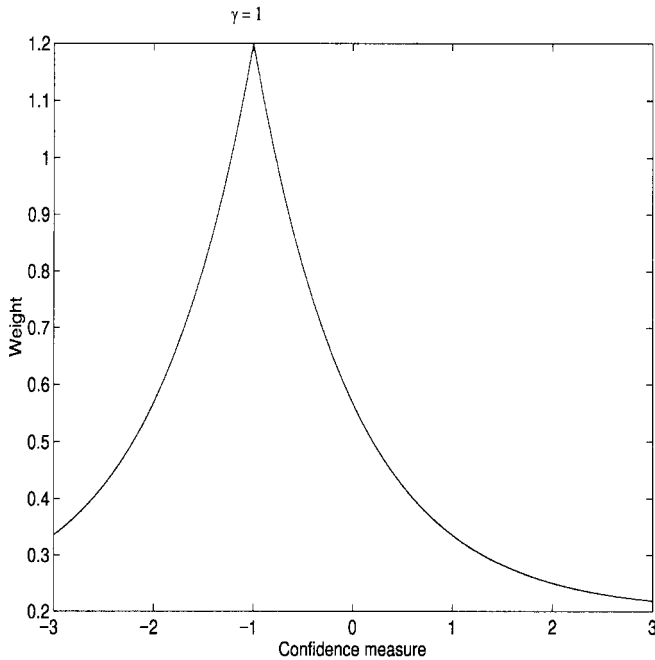


Fig. 2. Dependency of token weight on the confidence measure. The case where outliers are emphasized ( $\gamma = 1.0$ ).

degree of dissimilarity from the remaining word models. In this case, the likelihood ratio can be used in calculating the weights. First, we will show that weighting the training tokens does not violate the convergence property of maximum likelihood training. Consider the following auxiliary function (Baum and Eagon [11]):

$$Q(\lambda, \bar{\lambda}) = \frac{1}{P(\mathbf{X}|\lambda)} \sum_{\text{all } S} P(\mathbf{X}, S|\lambda) \log P(\mathbf{X}, S|\bar{\lambda}) \quad (1)$$

where  $\lambda$  and  $\bar{\lambda}$  are the previous and updated model parameters,  $\mathbf{X}$  is the observation sequence, and  $S$  is a particular state sequence through the HMM. Using this function, it can be shown that

$$Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda) \longrightarrow P(\mathbf{X}|\bar{\lambda}) \geq P(\mathbf{X}|\lambda). \quad (2)$$

Therefore, for a broad model class, the relation  $Q$  as a function of  $\bar{\lambda}$  will have a single critical point which is also the global maximum. So, if a new parameter model set  $\bar{\lambda}$  can be found which makes the right-hand side of the equation positive:

$$\log \frac{P(\mathbf{X}|\bar{\lambda})}{P(\mathbf{X}|\lambda)} \geq Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda) \quad (3)$$

then the model reestimation is guaranteed to improve  $P(\mathbf{X}|\lambda)$ . See [12] for a more complete discussion of the proof. In the above formulation a single observation is considered. The extension to  $N$  observations can be represented as

$$\frac{1}{N} \sum_{r=1}^N \log \frac{P(\mathbf{X}_r|\bar{\lambda})}{P(\mathbf{X}_r|\lambda)} \geq Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda). \quad (4)$$

Now, if we introduce a weight  $w_r$  on each training token, in other words modify the *a priori* probability of each training

token, the new inequality can be expressed as

$$\frac{1}{N} \sum_{r=1}^N \log \frac{w_r P(\mathbf{X}_r|\bar{\lambda})}{w_r P(\mathbf{X}_r|\lambda)} \geq Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda) \quad (5)$$

where  $w_r$  is the weight of  $r$ th training token. Since the weight terms in the numerator and denominator cancel each other, the inequality still holds. We therefore seek to establish a token weighting coefficient that emphasizes tokens based on the confidence that a token better represents the overall structure for the model parameters. The likelihood ratio, which is a measure of confidence on each token, can be formulated as follows on a frame-by-frame basis (i.e., frame sample  $n$ ):

$$C_{ikn} = \frac{P(\mathbf{X}_{ikn}|\lambda_i)}{\left( \prod_{j=1, j \neq i}^N P(\mathbf{X}_{ikn}|\lambda_j) \right)^{1/(N-1)}} \quad (6)$$

where  $\lambda_i$  is the  $i$ th word model,  $\mathbf{X}_{ikn}$  is the  $k$ th training token of the  $i$ th word across time index  $n$ , and  $N$  is the total number of words in the vocabulary. There are many ways to use this likelihood for a given speech application. For example, in large vocabulary continuous speech recognition (LVCSR) systems, an average value of  $C_{ikn}$  can be obtained over an extended window (e.g., ten frames) and used to obtain a weight measure for that portion of the speech signal. Naturally, to use this weight for LVCSR would require one to first find the probability of each subword model (i.e., diphone, triphone, etc.), employ a pruning threshold and find the top  $N$  models at each frame in the input speech sequence. This would allow us to weight the input speech training data differently across individual subword models. For the purposes of our study, and to simplify the notation, we will consider isolated word models with a value of  $C_{ik}$  obtained by averaging over a given word token observation value sequence

$$C_{ik} = \frac{1}{N_r} \sum_{n=1}^{N_r} C_{ikn}, \quad (7)$$

where  $N_r$  represents the number of frames in the observation sequence for the input token. With no loss of generality, we will employ  $C_{ik}$  as the likelihood ratio from this point, but suggest that alternate forms based on the desired field of view for the input Markov model can be easily incorporated. We also point out that here, we chose to keep the weight fixed for a particular token throughout the training process. It is entirely possible to allow the weight to change across training iterations, however this point does introduce additional issues regarding convergence properties that would have to be addressed. In terms of log-probabilities, the expression for the *confidence measure* can be written as

$$\ln(C_{ik}) = \ln(P(\mathbf{X}_{ik}|\lambda_i)) - \frac{1}{N-1} \sum_{j=1, j \neq i}^N \ln(P(\mathbf{X}_{ik}|\lambda_j)). \quad (8)$$

There are many possible ways to formulate a weight measure based on the above confidence measure. In our simulations, we employed the following weight expression:

$$w_{ik} = \alpha + \exp \left( - \left| \ln(P(\mathbf{X}_{ik} | \lambda_i)) - \left( \frac{1}{N-1} \sum_{j=1, j \neq i}^N \ln(P(\mathbf{X}_{ik} | \lambda_j))^{1/\nu} \right)^\nu + \gamma \right| \right) \quad (9)$$

which is similar to the misclassification measure used in [14]. By varying the value of  $\nu$ , one can take all the potential models into consideration. One extreme case is when  $\nu$  approaches  $\infty$ . For this case, the weight expression becomes

$$w_{ik} = \alpha + \exp(-|\ln(P(\mathbf{X}_{ik} | \lambda_i)) - \ln(P(\mathbf{X}_{ik} | \lambda_j)) + \gamma|) \quad (10)$$

where  $\lambda_j$  is the model with the largest likelihood given training token  $\mathbf{X}_{ik}$ . In the above equation,  $\alpha$  sets a floor on the minimum possible weight on each training token, and  $\gamma$  controls the level of outlier emphasis/deemphasis. In our experiments we used a value of 10 for  $\nu$ , and a value of 0.2 for  $\alpha$ . The value of  $\gamma$  depends on the application, and a positive value should be used where outlier emphasis is desired, and a negative value is suggested where outlier deemphasis is necessary. In Fig. 2, the dependency of the token weight on the confidence measure from (8) is plotted for the value of  $\gamma = 1$  in (9). In this figure, positive values of the confidence measure imply higher confidence on the training token. Even for the example in the figure, where outlier emphasis is desired, when the confidence is extremely low ( $< -1.0$ ) the weight starts decreasing. This was achieved by the use of the absolute value expression in order to prevent extreme outliers from changing the model drastically. A special case of selective HMM training is when no weight adjustment is performed (i.e., when  $\alpha = 1.0, \gamma = \pm\infty$ ), which corresponds to traditional HMM training. The forward-backward reestimation equations can then be adjusted to take into account the new set of weights for the training tokens as follows.

State transition matrix entries:

$$\begin{aligned} \bar{a}_{ij} &= \frac{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r-1} \gamma_n(i, j)}{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r-1} \sum_j \gamma_n(i, j)} \\ &= \frac{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r-1} \gamma_n(i, j)}{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r-1} \gamma_n(i)} \end{aligned} \quad (11)$$

Mixture coefficients:

$$\bar{c}_{jk} = \frac{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \xi_n(j, k)}{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \sum_{k=1}^L \xi_n(j, k)} \quad (12)$$

Mean vector entries:

$$\bar{\mu}_{jk} = \frac{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \xi_n(j, k) \bar{\mathbf{x}}_n}{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \xi_n(j, k)} \quad (13)$$

Covariance matrix entries:

$$\bar{\Sigma}_{jk} = \frac{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \xi_n(j, k) (\bar{\mathbf{x}}_n - \mu_{jk})(\bar{\mathbf{x}}_n - \mu_{jk})^t}{\sum_{r=1}^R \frac{w_r}{P_r} \sum_{n=1}^{N_r} \xi_n(j, k)} \quad (14)$$

where  $\xi_n(j, k)$  is the probability of being in state  $j$  at time frame  $n$  with the  $k$ th mixture component accounting for  $\bar{\mathbf{x}}_{rn}$ , i.e.,

$$\xi_n(j, k) = \left[ \frac{\alpha_n(j) \beta_n(j)}{\sum_{j=1}^M \alpha_n(j) \beta_n(j)} \right] \left[ \frac{c_{jk} f_{jk}(\bar{\mathbf{x}}_{rn})}{\sum_{m=1}^L c_{jm} f_{jm}(\bar{\mathbf{x}}_{rn})} \right] \quad (15)$$

In these equations,  $\gamma_n(i, j)$  represents the *a posteriori* probability of transition from state  $i$  to  $j$  at time  $n$ . The terms  $R, N_r$  represent the number of tokens, and the total number of frames for the  $r$ th training token, respectively. Finally,  $P_r$  is the probability of the  $r$ th training token given the model [i.e.,  $P(\mathbf{X}_r | \lambda)$ ]. The proposed selective training method discriminates among the available training data based on its match to its true class, and its dissimilarity from the remaining model classes. Also, the amount of discrimination can be controlled with a single parameter which makes the algorithm flexible for a number of applications.

The proposed method does not change the reestimation equations in a way that would disrupt convergence properties. It only adjusts the training set over which the maximization of the likelihood is performed. In this respect it differs from MMIE and other corrective training methods.

### III. EVALUATIONS

In order to evaluate the proposed selective training method, both simulated and actual speech data was employed. Simulated data was generated in order to emphasize the motivation behind the suggested approach. Moreover, simulated data also allows one to formulate good examples to show

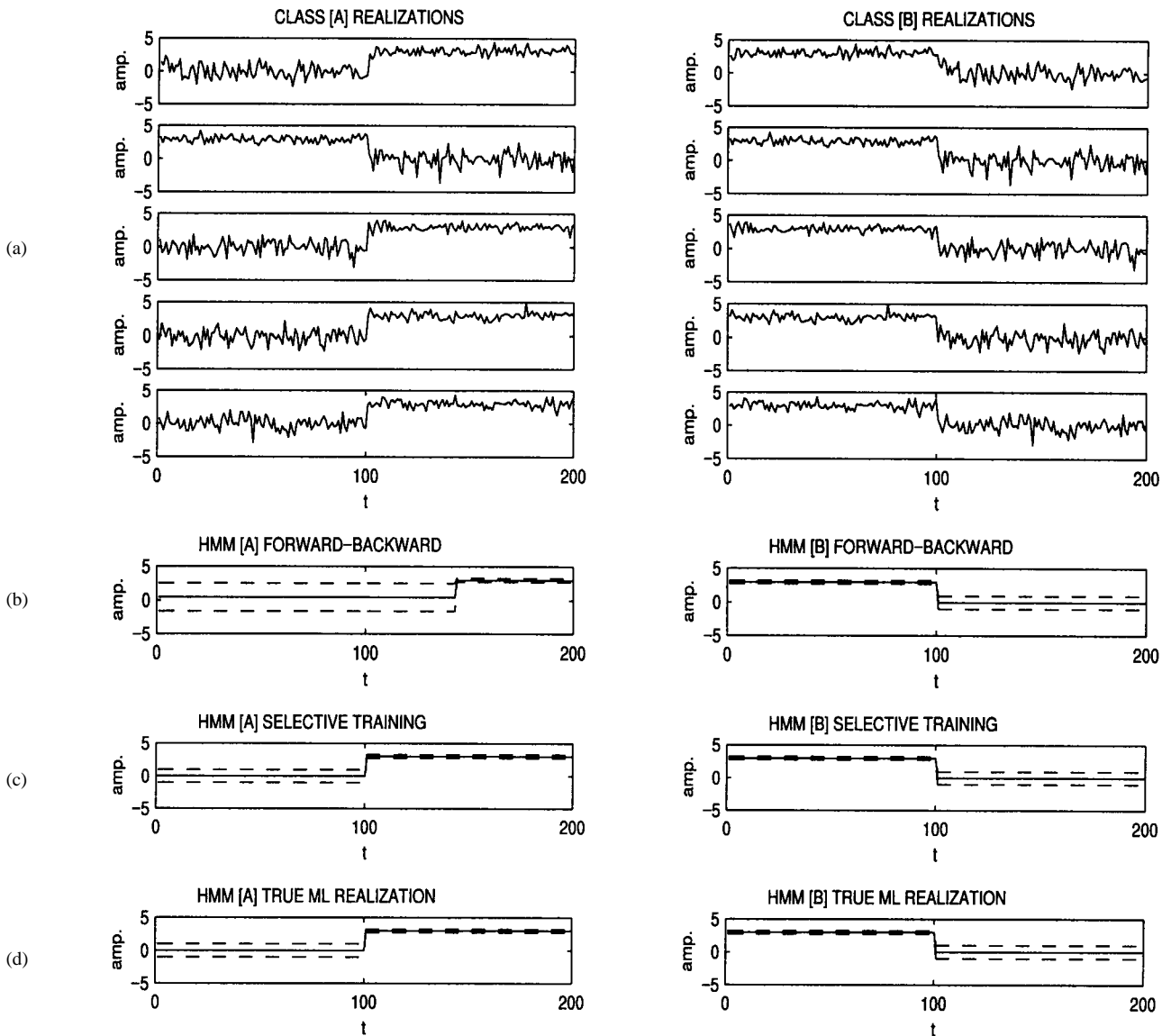


Fig. 3. Controlled simulation example for the selective training method.

the quantitative level of improvement that can be gained in modeling accuracy by using the proposed selective training method.

#### A. Experiment on Simulated Data

In this example, the application of selective training on simulated data is considered. A simulated data set was formed by generating five realizations of two types of data of 200 samples each [see Fig. 3(a)]. Class A realizations are statistically generated data sets from a Gaussian distribution with  $N(0, 1)$  (i.e., zero mean, unit variance) for the first 100 samples, and  $N(3, 0.25)$  for the second 100 samples. For Class B, the order of Gaussian distributions were reversed. Only one of the A type realizations (second from the top in the figure) was intentionally switched with a B type. After applying the forward-backward reestimation algorithm on Class A and B realizations, the following two-state models

were generated.

Class A HMM:

State transition matrix:

$$\begin{bmatrix} 0.9930 & 0.0070 \\ 0 & 1.0000 \end{bmatrix}$$

State 1 output probability distribution :

$N(0.46, 2.08)$

State 2 output probability distribution :

$N(3.01, 0.27)$

Class B HMM:

State transition matrix:

$$\begin{bmatrix} 0.9900 & 0.0100 \\ 0 & 1.0000 \end{bmatrix}$$

State 1 output probability distribution :

$N(2.99, 0.26)$

State 2 output probability distribution :

$N(-0.04, 0.97)$ .

The maximum likelihood (ML) model realizations for these two models are shown in the Fig. 3(b). The ML realizations are generated using the HMM parameters as follows: the self transition probabilities  $a_{ii}$  are used to determine the number of samples that correspond to each state [i.e.,  $(1/1 - a_{ii})$ ]. The means for each state are represented as solid lines and the dashed lines correspond to the variance associated with each state mean. It should be clear that the outlier in training set A [i.e., the second simulated data set in A, Fig. 3(a)] has caused the variance in state one to be extremely large, as well as marking the state transition much later in time than it normally would have occurred.

After employing the selective HMM training method, the new models are obtained as follows.

Class A HMM:

State transition matrix:

0.990	0.010
0	1.000

State 1 output probability distribution :  
N(-0.02,0.99)

State 2 output probability distribution :  
N(3.00,0.26)

Class B HMM:

State transition matrix:

0.990	0.010
0	1.000

State 1 output probability distribution :  
N(3.00,0.26)

State 2 output probability distribution :  
N(-0.04,0.97).

The ML model realizations after employing the selective training method are shown in Fig. 3(c). The true ML HMM realizations, using the original parameters that we used to generate class A and B realizations are shown in Fig. 3(d). The advantage of selective training is clearly illustrated by comparing the ML HMM realizations for Class A in Fig. 3(b)–(d). With standard forward-backward training, one can see the late transition from the first to second state as well as an incorrect shift in the mean vector and large variance associated with the first state. The late state transition and shifts in the first state mean and variance are corrected when selective training is employed. As HMM realizations for model B illustrate, when the training data is correctly labeled, selective training reduces to traditional ML model estimation. It also shows how effective the selective training method can be in recovering from labeling errors and providing more accurate models. However, it should be noted that this example employed idealized synthetic data in order to demonstrate the idea behind the selective training method.

### B. Selective Training for Accent Classification

The proposed method for selective training is next evaluated on an accent speech data base that was collected at the Robust

Speech Processing Laboratory, Duke University, in order to address the problem of accent classification [1], [2], [10]. A vocabulary of isolated words and phrases was established that contains accent sensitive phonemes or phoneme combinations [2], [10]. Vocabulary choice was based on a literature review of language education of American English as a second language. A portion of the speech corpus was collected using a head-mounted microphone in a quiet office environment, while the remaining part was collected through an on-line telephone interface at an 8 kHz sampling rate. All speakers were from the general Duke University community, with 43 speakers using microphone data entry, and 68 speakers using on-line telephone data entry. The test vocabulary consists of 20 accent-sensitive isolated words such as: *aluminum*, *thirty*, *bringing*, *target*, *bird*, and four American English phrases. Every speaker repeated each word five times. The accent speech data base includes neutral American English, and English under the following accents: German, Chinese, Turkish, French, Persian, Spanish, Italian, Hindi, Rumanian, Japanese, Greek, and others. The studies conducted here focused on American English speech from 76 speakers (the data collected through head-mounted microphone was bandpass filtered between 100–3800 Hz in order to provide spectral match with the data collected through the on-line telephone interface).

For the experiments conducted here, four accent types of American English that include neutral, Turkish, German, and Chinese were selected from the accent data base. For each word in the vocabulary, a continuous mixture HMM with two Gaussian mixtures per state was generated for each accent type. The number of states assigned to each word was proportional to the number of phonemes in the word. The number of states for the isolated vocabulary set typically ranged from seven to 21. All speech data was parameterized using mel-frequency cepstrum coefficients, delta mel-frequency cepstrum coefficients, energy, and delta energy. Nine speakers from each accent class were used for training, and a total of 40 speakers were set aside for open set testing. Using the traditional forward-backward training method, a total of 80 HMM's (20 words under four accent types) were trained. Overall, the total number of word tokens used in the training was 3250, and the total number of word tokens used for open testing was 3900. This baseline system resulted in an average classification rate of 56.3% for the open speaker set. The accent classification decisions were based on single words only. When the selective training method is employed with  $\gamma = -1.0$  in (9), the average classification rate improves to 58.7% (a 5.3% error reduction). The detailed results across accents are shown in Table I. It is interesting to note that the classification rate drops slightly for neutral accented American speakers, while it improves for all nonnative speakers. The improvement for the nonnative speaker set supports the fact that reducing the weight on slightly accented speakers for foreign accent models actually improves the accuracy of the accented word models. The drop in the neutral speaker set can be explained with the same reasoning: better modeling of the accented speech caused some test words of the native speakers who have similar acoustic characteristics to those accent types to be classified as accented speech. It is also worth mentioning that the closed

TABLE I  
ACCENT CLASSIFICATION PERFORMANCE IMPROVEMENT BY SELECTIVE TRAINING METHOD OVER STANDARD FORWARD-BACKWARD TRAINING

<i>Classification rates among 4 accents</i>		
<i>Test speakers</i>	<i>Forward-Backward</i>	<i>Selective Training</i>
Neutral	70.28%	68.87%
Turkish	54.51%	57.28%
Chinese	59.92%	63.45%
German	40.74%	45.18%
Overall	56.36%	58.70%

set accent classification rate actually dropped from 97.15% to 91.54% after the selective training method is applied. This result strongly indicates that deemphasizing the outliers in the training data improves the generated model's accuracy from a general perspective (i.e., reducing the impact of "over-training"), because open test results better reflect true system performance. When individual speaker results were analyzed, it was observed that selective training especially improved the classification rate among heavy accented speakers. This result is crucial, because a main goal in accent classification is to improve speech recognition accuracy by incorporating accent information. If a speaker exhibits a light accent, recognition performance will not degrade substantially. However, it is extremely important to be able to identify those speakers with strong accents, because it is these speakers that contribute most to speech recognition algorithm failure.

### C. Selective Training for Language Identification

Next, the selective training method is evaluated on the problem of language identification (ID). The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [19] was used in the experiment. Each message in the corpus was spoken by a unique speaker over a telephone channel and includes responses to ten prompts. Four of the prompts expect fixed utterance responses from speakers (e.g., "Please recite the seven days of the week," "Please say the numbers zero through ten"), and six of the prompts assume free responses (e.g., "Describe the climate in your home town," "Speak about any topic of your choice"). Altogether, the ten responses contained in each session comprise about 2 min of speech. Table II contains a listing of the number of messages per language in each of the two segments of the corpus: initial training, and development test sets.<sup>1</sup> In our evaluations, the initial training set (about 50 speakers per language) was used in training, and the development test set (about 20 speakers per language) was used for testing. Test utterances were extracted from the development test set according to the April 1993 National Institute of Standards and Technology (NIST) specification [15]:

"45 s" Utterance Testing: Language ID is performed on a set of 45 s utterances spoken by the development test speakers. These utterances are the first 45 s of the responses to the prompt "speak about any topic of your choice." OGI refers

TABLE II  
NUMBER OF SPEAKERS FROM EACH LANGUAGE GROUP FOR THE INITIAL TRAINING AND DEVELOPMENT TEST SETS IN OGI MULTILANGUAGE DATA BASE

<i>OGI MULTI-LANGUAGE SPEECH CORPUS</i>				
	<i>Initial Training</i>		<i>Development Test</i>	
<i>Language</i>	<i>Male</i>	<i>Female</i>	<i>Male</i>	<i>Female</i>
English	33	17	14	6
Farsi	39	10	15	4
French	40	10	15	5
German	25	25	11	9
Korean	47	3	13	4
Japanese	32	17	18	2
Mandarin	34	15	14	6
Spanish	34	16	16	4
Tamil	43	7	17	3
Vietnamese	31	19	16	4

to these utterances as "stories before the tone," and they are denoted *story-bt*.<sup>2</sup>

"10 s" Utterance Testing: Language ID is performed on a set of 10 s cuts from the same story utterances used in "45 s" testing. In the evaluations performed here only the 45 s utterances were tested.

In the language ID system, a Gaussian mixture model with 64 mixtures is employed for each language. While selective training has been proposed for hidden Markov models, here it is used in the context of a single Markov state with multiple mixtures. The feature set used in the system comprised eight cepstrum coefficients, eight delta cepstrum coefficients, and delta energy. The cepstrum coefficients are computed based on a new accent sensitive scale described in Arslan and Hansen [1], instead of the commonly used mel-scale, which was shown to result in improved performance for language ID. Language ID is different from accent identification in the sense that perfect labeling is possible for language ID whereas it is not possible for accent classification.<sup>3</sup> In the case of perfect labeling, one can emphasize the error tokens in the training in order to achieve minimum number of classification errors. In order to achieve this, a positive value for the parameter  $\gamma$  in (9) is assigned (i.e.,  $\gamma = +1.0$ ). The recognition rates before and after using the selective training method are shown in Table III. The overall error rate was reduced from 15.7 to 11.5% (a 27% improvement) after using the selective training method instead of standard forward-backward training.

### D. Selective Training for the E-Set

A final set of experiments was conducted to evaluate the performance of the proposed selective training method. These experiments involved the recognition of the highly confusable English E-set alphabet, namely, *b, c, d, e, g, p, t, v*, and *z*. In this evaluation, the Oregon Graduate Institute (OGI) Spelled

<sup>2</sup>A tone signaled the speaker when 45 s of speech had been collected indicating 15 s remaining.

<sup>3</sup>For language classification, a speaker makes a hard decision to speak one language. For accent classification, speakers will exhibit varying degrees of accent from a second language.

<sup>1</sup>The extended training and final test sets are not considered in this study.

TABLE III  
LANGUAGE ID PERFORMANCE IMPROVEMENT AFTER SELECTIVE TRAINING FOR PAIRWISE ENGLISH—OTHER EXPERIMENTS

<i>Error rates in classification of English versus other languages in OGI multi-language database</i>				
<i>Test set</i>	Closed Set		Open Set	
	<i>F-B Training</i>	<i>Selective Training</i>	<i>F-B Training</i>	<i>Selective Training</i>
English-Farsi	7.78	0.00	16.22	8.11
English-French	9.89	2.20	15.79	15.79
English-German	10.00	2.27	29.73	16.22
English-Japanese	4.35	1.09	8.57	11.43
English-Spanish	9.78	7.61	21.62	13.51
English-Korean	4.65	0.00	8.33	8.33
English-Mandarin	0.00	0.00	8.11	5.41
English-Tamil	3.45	1.15	16.22	8.11
English-Vietnam.	2.38	0.00	16.67	16.67
Overall	5.81	1.59	15.70	11.51

TABLE IV  
PERFORMANCE IMPROVEMENT IN THE RECOGNITION OF THE *E*-SET ALPHABET AFTER SELECTIVE TRAINING. SHOWN ARE CLOSED AND OPEN SPEAKER RESULTS USING BOTH THE FORWARD-BACKWARD AND SELECTIVE TRAINING METHODS

<i>Error rates in the E-set alphabet on OGI Spelled and Spoken Corpus</i>				
<i>Test set</i>	Closed Set		Open Set	
	<i>F-B Training</i>	<i>Selective Training</i>	<i>F-B Training</i>	<i>Selective Training</i>
<i>E-set</i>	9.1	6.7	33.3	30.1

and Spoken Word Telephone Corpus<sup>4</sup> was used as the test data base. The hand-labeled portion of the data base was used which consists of 100 speaker utterances of the letters of the alphabet. Seventy-six speakers were used in the training of the *E*-set alphabet, and the remaining 24 speakers were set aside for open testing. The feature set used in the system comprised eight cepstrum coefficients, eight delta cepstrum coefficients, energy, and delta energy. The HMM topology used in the system was a six-state left-to-right model with four mixtures in each state. The closed-set and open-set error rates before and after employing the selective training method are listed in Table IV. An error rate reduction of 26.4% is achieved in the closed set, and a 9.7% error rate reduction is achieved in the open set, which is a measurable improvement over traditional forward-backward training.

#### IV. SUMMARY AND CONCLUSIONS

A new training procedure has been proposed in order to address the adverse influence of outliers in HMM classification systems. The proposed selective training method adjusts the weights of the training tokens in the reestimation equation formulation in order to control the influence of outliers in the training data. Here, we chose to keep the weight fixed for a particular token throughout the training process, though it is possible, if the convergence properties are addressed, to allow the weight to change across training iterations. When the proposed method was applied to the problem of accent

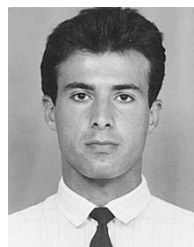
classification among four accent classes, an improvement of 5.3% was achieved in the error rate. The proposed method was next applied to the language ID problem, which resulted in a 27% error rate improvement for the English–other pairwise experiments. Finally, the proposed method was applied to recognition of the *E*-set alphabet, and a 9.7% reduction in error rate was achieved. The algorithm is flexible enough to be used for a number of different classification problems. Basically, there are two modes of operation for the algorithm that involve either outlier emphasis or outlier deemphasis. For the problem of accent classification, outlier deemphasis proved to be useful, since tokens from slightly accented speakers were not weighted as much as tokens from heavy accented speakers, whereas for language identification and *E*-set recognition tasks, outlier emphasis resulted in improved performance. In each case, a measurable improvement over the traditional methods is achieved, which shows the strength of the proposed algorithm.

It should be noted that the improvement in model characterization for speech applications in classification and recognition are achieved with an incremental increase in the complexity of the existing system, requiring only a few extra iterations in the forward-backward algorithm. This represents an advantage in using selective training for many applications in speech recognition where computational complexity and speed of operation may be a factor. While a number of applications were considered in this study, selective training could also be employed as a general approach for updating existing models in speech recognition, speaker classification, or other speech problems.

<sup>4</sup>The OGI Spelled and Spoken Word Telephone Corpus is available through the Linguistic Data Consortium (LDC) at <http://www ldc.upenn.edu/>.

## REFERENCES

- [1] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, pp. 28–40, July 1997.
- [2] ———, "Language accent classification in American English," *Speech Commun.*, vol. 18, pp. 353–368, Aug. 1996.
- [3] X. Aubert, R. Haeb-Umbach, and H. Ney, "Continuous mixture densities and linear discriminant analysis for improved context dependent acoustic models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Minneapolis, MN, 1993, pp. II:648–651.
- [4] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tokyo, Japan, 1986, pp. 49–52.
- [5] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe, "Global optimization of a neural network—Hidden Markov model hybrid," *IEEE Trans. Neural Networks*, vol. 3, pp. 252–259, Mar. 1992.
- [6] C. M. Ayer *et al.*, "A discriminatively derived linear transform for improved speech recognition," in *Proc. Eurospeech*, Berlin, Germany, 1993.
- [7] R. A. Fisher, "The use of multiple measures in taxonomic problems," *Contr. Math. Stat.*, pp. 32.179–32.188, 1950.
- [8] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "Generalization of the Baum algorithm to rational objective functions," *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 631–634, 1989.
- [9] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. I, pp. 13–16, 1992.
- [10] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 836–839, 1995.
- [11] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp. 360–363, 1967.
- [12] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh, U.K.: Edinburgh Univ. Press, 1990.
- [13] M. J. Hunt, S. M. Richardson, D. C. Bateman, and A. Piau, "An investigation of PLP and IMELDA acoustic representations and of their potential for combination," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 881–884, 1991.
- [14] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, 1992.
- [15] A. F. Martin, "Language ID guidelines and results," Tech. Rep., Spoken Lang. Processing Group., Nat. Inst. Stand. Technol., Gaithersburg, MD, 1993.
- [16] B. Merialdo, "Phonetic recognition using hidden Markov models and maximum mutual information training," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1988, pp. 111–114.
- [17] D. P. Morgan and C. L. Scofield, *Neural Networks and Speech Processing*. Boston, MA: Kluwer, 1991.
- [18] N. Morgan and H. Bourlard, "Continuous speech recognition: An approach to the hybrid HMM/connectionist approach," *IEEE Signal Processing Mag.*, vol. 12, pp. 25–42, May, 1995.
- [19] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. Inter. Conf. Spoken Lang. Proc.*, 1992, vol. 2, pp. 895–898.
- [20] Y. Normandin, R. Cardin, and R. De Mori, "On the nature of the foreign accent syndrome: A case study," *Brain Lang.*, vol. 31, pp. 215–244, 1987.
- [21] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 449–452.
- [22] E. S. Parris and M. J. Carey, "Estimating linear discriminant parameters for continuous density hidden Markov models," in *Proc. Int. Conf. Spoken Lang. Processing*, Yokohama, Japan, 1994.
- [23] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 4–16, 1986.



**Levent M. Arslan** (M'98) was born in Besni, Turkey on September 2, 1968. He received the B.S. degree in electrical engineering from Boğazici University, Istanbul, Turkey in 1991, and the M.S. and Ph.D. degrees in electrical engineering from Duke University, Durham, NC, in 1993 and 1996, respectively.

During the summers of 1994 and 1995 he was a Visiting Speech Researcher at Texas Instruments, Dallas, TX. From 1996 to 1998, he was with the Entropic Research Laboratory, Washington, DC, as a Member of Technical Staff. In the fall of 1998, he joined the Electrical and Electronics Engineering Department, Boğazici University, as an Assistant Professor. His research interests include digital signal processing, speech enhancement, speech analysis, speech synthesis, voice conversion, speech recognition in noisy environments, and 3-D face synthesis. He has a number of journal papers and several patents in these areas.



**John H. L. Hansen** (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982. He received the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

In 1988, he joined the faculty of Duke University, Durham, NC, Department of Electrical Engineering, where he established and directed the Robust Speech Processing Laboratory (RSPL). He also received a secondary appointment in the Department of Biomedical Engineering. Prior to joining the Duke faculty, he was employed by the RCA Solid State Division, Somerville, NJ (1981–1982), and Dranetz Engineering Laboratories, Edison, NJ (1978–1981). In January 1999, he moved RSPL to the University of Colorado, Boulder, CO, where along with R. Cole and W. Ward, established a new Center for Spoken Language Understanding. He has served as a technical consultant to industry and the U.S. government, including AT&T Bell Laboratories, IBM, VeriVoice, DOD, and Sparta, in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. His research interests span the areas of digital signal processing, analysis and modeling of speech under stress and/or pathology, speech enhancement and feature estimation in noise, robust speech recognition with current speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust speech feature enhancement for voice communications, and source generator based speech modeling for robust recognition in stress, noise, and Lombard effect. He is the author of numerous journal and conference papers in the field of speech processing and communications, and is co-author of the textbook *Discrete-Time Processing of Speech Signals* (Englewood Cliffs, NJ: Prentice-Hall, 1993).

Dr. Hansen was an invited tutorial speaker for ICASSP'95 and the ESCA-NATO Speech Under Stress Research Workshop (Lisbon, Portugal). He has served as Chairman for the IEEE Communications and Signal Processing Society of North Carolina (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), and Tutorial Chair for ICASSP'96. He served as an Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1993–1998), and is currently an Associate Editor for IEEE SIGNAL PROCESSING LETTERS. He has also served as Guest Editor of the October 1994 special issue on Robust Speech Recognition of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award in 1993, a National Science Foundation's Research Initiation Award in 1990, and has been named a Lilly Foundation Teaching Fellow.