Robust Speaker Recognition in Noisy Conditions

Ji Ming, *Member, IEEE*, Timothy J. Hazen, *Member, IEEE*, James R. Glass, *Senior Member, IEEE*, and Douglas A. Reynolds, *Senior Member, IEEE*

***Abstract*—This paper investigates the problem of speaker identification and verification in noisy conditions, assuming that speech signals are corrupted by environmental noise, but knowledge about the noise characteristics is not available. This research is motivated in part by the potential application of speaker recognition technologies on handheld devices or the Internet. While the technologies promise an additional biometric layer of security to protect the user, the practical implementation of such systems faces many challenges. One of these is environmental noise. Due to the mobile nature of such systems, the noise sources can be highly time-varying and potentially unknown. This raises the requirement for noise robustness in the absence of information about the noise. This paper describes a method that combines multicondition model training and missing-feature theory to model noise with unknown temporal-spectral characteristics. Multicondition training is conducted using simulated noisy data with limited noise variation, providing a "coarse" compensation for the noise, and missing-feature theory is applied to refine the compensation by ignoring noise variation outside the given training conditions, thereby reducing the training and testing mismatch. This paper is focused on several issues relating to the implementation of the new model for real-world applications. These include the generation of multicondition training data to model noisy speech, the combination of different training data to optimize the recognition performance, and the reduction of the model's complexity. The new algorithm was tested using two databases with simulated and realistic noisy speech data. The first database is a redevelopment of the TIMIT database by rerecording the data in the presence of various noise types, used to test the model for speaker identification with a focus on the varieties of noise. The second database is a handheld-device database collected in realistic noisy conditions, used to further validate the model for real-world speaker verification. The new model is compared to baseline systems and is found to achieve lower error rates.**

***Index Terms*—Missing-feature theory, multicondition training, noise compensation, noise modeling, speaker recognition.**

I. Introduction

ACCURATE speaker recognition is difficult due to a number of factors, with handset/channel mismatch and environmental noise being two of the most prominent. Recently, much research has been conducted with a focus on reducing the effect of handset/channel mismatch. Linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains. Some of the techniques were first developed in speech recognition research. Examples of the feature compensation methods include well-known filtering techniques such as cepstral mean subtraction or RASTA (e.g., [1]–[5]), discriminative feature design (e.g., [6]–[9]), and various feature transformation methods such as affine transformation, nonlinear spectral magnitude normalization, feature warping, and short-time Gaussianization (e.g., [10]–[13]). Score-domain compensation aims to remove handset-dependent biases from the likelihood ratio scores. The most prevalent methods include H-norm [14], Z-norm [15], and T-norm [16]. Examples of the model-domain compensation methods include the speaker-independent variance transformation [17], and the transformation for synthesizing supplementary speaker models for other channel types from multichannel training data [18]. Additionally, channel mismatch has also been dealt with by using model adaptation methods, which effectively use new data to learn channel characteristics (e.g., [19], [20]).

To date, research has targeted the impact of environmental noise through filtering techniques such as spectral subtraction or Kalman filtering [21], [22], assuming *a priori* knowledge of the noise spectrum. Other techniques focus on noise compensation, for example, parallel model combination (PMC) [23]–[25], or Jacobian environmental adaptation [26], [27], assuming the availability of a statistical model of the noise or environment. Researchers in [28] and [29] have discussed the use of microphone arrays to improve noise robustness. Recent studies on missing-feature approaches suggest that, when knowledge of the noise is insufficient for cleaning up the speech data, one may alternatively ignore the severely corrupted speech data and base the recognition only on the data with little or no contamination (e.g., [30], [31]). Missing-feature techniques are effective given partial noise corruption, a condition that may not be realistically assumed for many real-world problems.

This paper investigates the problem of speaker recognition using speech samples distorted by environmental noise. We assume a highly unfavorable scenario: an accurate estimation of the nature and characteristics of the noise is difficult, if not impossible. As such, traditional techniques for noise removal or compensation, which usually assume a prior knowledge of the noise, become inapplicable. It is likely that the adoption of this worst-case scenario will be necessary in many real-world applications, for example, speaker recognition over handheld devices or the Internet. While the technologies promise an additional biometric layer of security to protect the user, the practical implementation of such systems faces many challenges. For example, a handheld-device based recognition system needs

Manuscript received November 28, 2005; revised January 28, 2007. This work was supported in part by Intel Corporation, the Queen's University Belfast Exchange Scheme, and the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

J. Ming is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: j.ming@qub.ac.uk).

T. J. Hazen and D. A. Reynolds are with the MIT Lincoln Laboratory, Lexington, MA 02420 USA (e-mail: hazen@csail.mit.edu; dar@ll.mit.edu).

J. R. Glass is with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA (e-mail: glass@mit.edu).

Digital Object Identifier 10.1109/TASL.2007.899278

to be robust to noisy environments, such as office/street/car environments, which are subject to unpredictable and potentially unknown sources of noise (e.g., abrupt noises, other-speaker interference, dynamic environmental change, etc.). This raises the need for a method that enables the modeling of unknown, time-varying noise corruption without assuming prior knowledge of the noise statistics. This paper describes such a method. The new approach is an extension of missing-feature theory, i.e., recognition based only on reliable data but robust to any corruption type, including full corruption that affects all time-frequency components of the speech. This is achieved by a combination of multicondition model training and missing-feature theory. Multicondition training provides a "coarse" compensation for the noise; missing-feature theory is applied to deal with the remaining training and testing mismatch, by ignoring noise variation outside the given training conditions. The paper demonstrates that based on limited training data, the new approach has the potential to model a wide variety of noise conditions without assuming specific information about the noise.

As preliminary studies, the proposed approach was first tested for speech recognition (e.g., [32]) and later for speaker identification [33], both using artificially synthesized noisy speech data. This paper extends the previous research by focusing on several issues relating to the implementation of the new approach towards real-world applications. Specifically, we will study new methods for generating multicondition training data to better characterize real-world noisy speech, investigate the combination of training data of different characteristics to optimize the recognition performance, and look into the reduction of the model's complexity through a balance with the model's noise-condition resolution. The proposed model was evaluated using two databases with simulated and realistic noisy speech data. The first database is a redevelopment of the TIMIT database by rerecording the data in various controlled noise conditions, with a focus on the varieties of noise. The proposed model, along with the methods for generating the training data and reducing the model complexity, was tested and developed on this database for speaker identification. The second database is a handheld-device database collected in realistic noisy conditions. The new model was tested on this database for speaker verification assuming limited enrollment data. This study serves as a further validation of the proposed model by testing on a real-world application.

The remainder of this paper is organized as follows. Section II describes the new model and the methods for generating the training data and controlling the model's complexity. Section III presents the experimental results for speaker identification on the noisy TIMIT database, and Section IV presents the experimental results for speaker verification on the realistic handheld-device database. Finally, Section V presents a summary of the paper.

II. Proposed Method

A. Model

Let Φ_0 denote the training data set, containing *clean* speech data, for speaker S, and let $p(X|S,\Phi_0)$ represent the likelihood function of frame feature vector X associated with speaker S trained on data set Φ_0. In this paper, we assume that each frame vector X consists of N subband features: $X = (x_1, x_2, \ldots, x_N)$, where x_n represents the feature for the nth subband. We obtain X by dividing the whole speech frequency-band into N subbands, and then calculating the feature coefficients for each subband independently of the other subbands. The subband feature framework has been used in speech recognition (e.g., [34] and [35]) for isolating local frequency-band corruption from spreading into the features of the other bands.

The proposed approach for modeling noise includes two steps. The first step is to generate multiple copies of training set Φ_0, by introducing corruption of different characteristics into Φ_0. Primarily, we could add white noise at various signal-to-noise ratios (SNRs) to the clean training data to simulate the corruption. Assume that this leads to augmented training sets $\Phi_0, \Phi_1, \ldots, \Phi_L$, where Φ_l denotes the lth training set derived from Φ_0 with the inclusion of a certain noise condition. Then, new likelihood function for the test frame vector can be formed by combining the likelihood functions trained on the individual training sets

$$p(X|S) = \sum_{l=0}^{L} p(X|S,\Phi_l) P(\Phi_l|S) \tag{1}$$

where $p(X|S,\Phi_l)$ is the likelihood function of frame vector X trained on set Φ_l, and $P(\Phi_l|S)$ is the prior probability for the occurrence of the noise condition Φ_l, for speaker S. Equation (1) is a multicondition model. A recognition system based on (1) should have improved robustness to the noise conditions seen in the training sets Φ_l, as compared to a system based on $p(X|S,\Phi_0)$.

The second step of the new approach is to make (1) robust to noise conditions not fully matched by the training sets Φ_l without assuming extra noise information. One way to this is to ignore the heavily mismatched subbands and focus the score only on the matching subbands. Let $X = (x_1, x_2, \ldots, x_N)$ be a test frame vector and $X_l \subset X$ be a subset in X containing all the subband features corrupted at noise condition Φ_l. Then, using X_l in place of X as the test vector for each training noise condition, (1) can be redefined as

$$p(X|S) = \sum_{l=0}^{L} p(X_l|S,\Phi_l) P(\Phi_l|S) \tag{2}$$

where $p(X_l|S,\Phi_l)$ is the marginal likelihood of the matching feature subset X_l, derived from $p(X|S,\Phi_l)$ with the mismatched subband features ignored to improve mismatch robustness between the test frame X and the training noise condition Φ_l. For simplicity, assume independence between the subband features. So the marginal likelihood $p(X_{\text{sub}}|S,\Phi_l)$ for any subset $X_{\text{sub}} \subset X$ can be written as

$$p(X_{\text{sub}}|S,\Phi_l) = \prod_{x_n \in X_{\text{sub}}} p(x_n|S,\Phi_l) \tag{3}$$

where $p(x_n|S,\Phi_l)$ is the likelihood function of the nth subband feature for speaker S trained under noise condition Φ_l.

Multicondition or multistyle model training [e.g., (1)] has been a common method used in speech recognition (e.g., [36] and [37]), to account for varying noise sources or speaking styles. The new model expressed in (2) is novel in that it combines multicondition model training with missing-feature theory, to ignore noise variation outside the given training conditions. This combination makes it possible to account for a wide variety of testing conditions based on limited training conditions, as will be demonstrated later in the experiments.

We say that missing-feature theory is applied in (2) for ignoring the mismatched subband features. However, it should be noted that the approach expressed in (2) extends beyond traditional missing-feature approaches in one aspect: traditional approaches assess the usability of a feature against its clean data, while the new approach assesses this against the data containing variable degrees of corruption, modeled by the different training conditions Φ_0 through Φ_L. This allows the model to use noisy features, close to or matched by the noisy training conditions, for recognition. These noisy features, however, may become less usable or unusable with traditional missing-feature approaches due to their mismatch against the clean data.

Given a test frame X, the matching feature subset X_l for each training noise Φ_l may be defined as the subset in X that gains maximum likelihood over the appropriate noise condition. Such an estimate for X_l is not directly obtainable from (3) by maximizing $p(X_{\rm sub}|S,\Phi_l)$ with respect to $X_{\rm sub}$. This is because the values of $p(X_{\rm sub}|S,\Phi_l)$ for different sized subsets $X_{\rm sub}$ are of a different order of magnitude and are thus not directly comparable. One way around this is to select the matching feature subset X_l for noise condition Φ_l that produces maximum likelihood for noise condition Φ_l, *as compared to* the likelihoods of the same subset produced for the other noise conditions $\Phi_{l'} \neq \Phi_l$, for each speaker S. This effectively leads to a posterior probability formulation of (2). Define the posterior probability of speaker S and noise condition Φ_l given test subset $X_{\rm sub}$ as

$$P(S,\Phi_l|X_{\rm sub}) = \frac{p(X_{\rm sub}|S,\Phi_l)P(S,\Phi_l)}{\sum_{S',l'} p(X_{\rm sub}|S',\Phi_{l'})P(S',\Phi_{l'})}. \quad (4)$$

On the right, (4) performs a normalization for $p(X_{\rm sub}|S,\Phi_l)$ using the average likelihood of subset $X_{\rm sub}$ calculated over all speakers and training noise conditions, with $P(S,\Phi_l) = P(\Phi_l|S)P(S)$ being a prior probability of speaker S and noise condition Φ_l. Maximizing posterior probability $P(S,\Phi_l|X_{\rm sub})$ with respect to $X_{\rm sub}$ leads to an estimate for the matching feature subset X_l that effectively maximizes the likelihood ratios $p(X_l|S,\Phi_l)/p(X_l|S',\Phi_{l'})$ for (S,Φ_l) compared to all $(S',\Phi_{l'}) \neq (S,\Phi_l)$.[1]

[1]Dividing the numerator and denominator of (4) by $p(X_{\rm sub}|S,\Phi_l)$ gives the equation shown at the bottom of the page. Therefore, maximizing $P(S,\Phi_l|X_{\rm sub})$ with respect to $X_{\rm sub}$ is equivalent to the maximization of the likelihood ratios $p(X_{\rm sub}|S,\Phi_l)/p(X_{\rm sub}|S',\Phi_{l'})$ by choosing $X_{\rm sub}$.

To incorporate the posterior probability (4) into the model, we first rewrite (1) in terms of $P(S,\Phi_l|X)$, i.e., the posterior probabilities of speaker S and noise condition Φ_l given frame vector X. Using Bayes's rule it follows

$$p(X|S) = \frac{P(S|X)p(X)}{P(S)} = \frac{\sum_{l=0}^{L} P(S,\Phi_l|X)}{P(S)}p(X). \quad (5)$$

The last term in (5), $p(X)$, is not a function of the speaker index and thus has no effect in recognition. Replacing $P(S,\Phi_l|X)$ in (5) with the optimized posterior probability for the test feature subset and assuming an equal prior $P(S)$ for all the speakers, we obtain an operational version of (2) for recognition

$$p(X|S) \propto \sum_{l=0}^{L} \max_{X_{\rm sub}\subset X} P(S,\Phi_l|X_{\rm sub}) \quad (6)$$

where $P(S,\Phi_l|X_{\rm sub})$ is defined in (4) with $P(S,\Phi_l)$ replaced by $P(\Phi_l|S)$ due to the assumption of a uniform $P(S)$.

The search in (6) for the matching feature subset can be computationally expensive for large frame vectors X. We can simplify the computation by approximating each $p(X_{\rm sub}|S,\Phi_l)$ in (4) using the probability for the union of all subsets of the same size as $X_{\rm sub}$. As such, $p(X_{\rm sub}|S,\Phi_l)$ can be written, with the size of $X_{\rm sub}$ indicated in brackets, as [38]

$$p\left(X_{\rm sub}(M)|S,\Phi_l\right) \propto \sum_{{\rm all}X'_{\rm sub}(M)\subset X} p\left(X'_{\rm sub}(M)|S,\Phi_l\right) \quad (7)$$

where $X_{\rm sub}(M)$ represents a subset with M features ($M \leq N$). Since the sum in (7) includes all feature subsets, it includes the matching feature subset that can be assumed to dominate the sum due to the best data-model match. Therefore, (4) can be rewritten, by replacing $p(X_{\rm sub}|S,\Phi_l)$ with $p(X_{\rm sub}(M)|S,\Phi_l)$, as

$$P\left(S,\Phi_l|X_{\rm sub}(M)\right) = \frac{p(X_{\rm sub}(M)|S,\Phi_l)P(S,\Phi_l)}{\sum_{S',l'} p(X_{\rm sub}(M)|S',\Phi_{l'})P(S',\Phi_{l'})}. \quad (8)$$

Note that (8) is not a function of the identity of $X_{\rm sub}$ but only a function of the size of $X_{\rm sub}$ (i.e., M). Using $P(S,\Phi_l|X_{\rm sub}(M))$ in place of $P(S,\Phi_l|X_{\rm sub})$ in (6), we therefore effectively turn the maximization for the exact matching feature subset $\max_{X_{\rm sub}\subset X} P(S,\Phi_l|X_{\rm sub})$, of a complexity of $O(2^N)$ to the maximization for the size of the matching feature subset $\max_M P(S,\Phi_l|X_{\rm sub}(M))$ with a lower complexity of $O(N)$. The sum in (7) over all $p(X_{\rm sub}(M)|S,\Phi_l)$ for a given number of M features, for $0 < M \leq N$, can be computed efficiently using a recursive algorithm assuming independence between the subbands [i.e., (3)]. We call (8) the *posterior union model* (PUM), which has been studied previously (e.g., [39]) as a missing-feature approach without requiring identity of the

$$P(S,\Phi_l|X_{\rm sub}) = \frac{P(S,\Phi_l)}{P(S,\Phi_l) + \sum_{(S',\Phi_{l'})\neq(S,\Phi_l)} P(S',\Phi_{l'})p(X_{\rm sub}|S',\Phi_{l'})/p(X_{\rm sub}|S,\Phi_l)}$$

noisy data. The new model (6) is reduced to a PUM with single, clean condition training (i.e., $L = 0$).

So far we have discussed the calculation of the likelihood for a single frame. The likelihood of a speaker given an utterance with T frames $X_1^T = \{X_1, X_2, \ldots, X_T\}$ can be defined as

$$p\left(X_1^T|S\right) = \left[\prod_{t=1}^{T} p(X_t|S)\right]^{1/T} \tag{9}$$

where $p(X_t|S)$ is defined by (6). Since $p(X_t|S)$ is a properly normalized probability measure, the value of $p(X_1^T|S)$, with normalization against the length of the utterance as shown in (9), is used directly for speaker verification as well as for speaker identification in our experimental studies.

B. Training Data Generation and Model Complexity Reduction

As shown in (2), the new model effectively practices a reconstruction of the test noise condition using a limited number of training noise conditions. To make the model suitable for a wide variety of noises, the multicondition training sets $\Phi_1, \ldots, \Phi_L$ may be created from Φ_0 (i.e., the clean training set) by adding white noise to the clean training data at consecutive SNRs, with each Φ_l corresponding to a specific SNR. This accounts for the noise over the full frequency range and a wide amplitude range and therefore allows the expression of sophisticated noise spectral structures by piecewise (i.e., bandwise) approximation. Instead of white noise, we may also consider the use of low-pass filtered white noise at various SNRs in the creation of the multicondition training data. The low-pass filtering simulates the high-frequency rolloff characteristics seen in many microphones. Finally, a combination of different types of noise, including real noise data as in common multicondition model training, can be used to create the training data for the model. A simple example of the combination will be demonstrated in the paper. Without prior knowledge of the structure of the test noise, a uniform noise-condition prior $P(\Phi_l|S)$ can be used to combine different noise conditions.

In the above, we assume that the noisy training data are generated by adding noises electronically to the clean training data. The potential of the new model, that allows the use of a limited number of noise conditions to model potentially arbitrary noise conditions, makes it feasible to add noise acoustically into the training data, thereby more closely matching the physical process of how real-world noisy test data are generated. Fig. 1 shows an example, in which white noise at various SNRs are added *acoustically* to clean speech to produce the multicondition noisy training data. The new system shares the same principle as the systems used to collect HTIMIT [40], NTIMIT [43], and CTIMIT [42], which were attempting to model handset, telephone line, and cellular channel noise by rerecording the TIMIT sentences after transmission over the appropriate handsets or networks. The new system is designed to generate training data for the new model, with an attempt to model general environmental noise. In the system shown, loudspeakers are used to simultaneously play clean speech recordings and wide-band noise at different controlled volumes (to simulate white noise of different SNRs), and microphones are used to collect the mixed data that are used to train the new

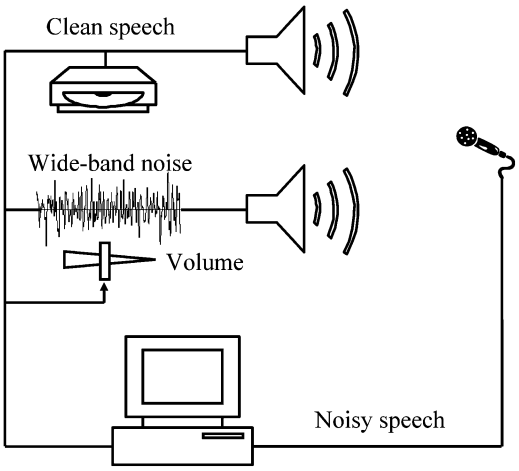


Fig. 1. Illustration of the system used to generate multicondition training data for the new model, with wide-band noise of different volumes added acoustically to the clean training data. This system is also used in the experiments to produce noisy test data, by replacing the wide-band noise source with a test noise source.

model. This is considered to be feasible because in this data collection we only need to consider a limited number of noise conditions, e.g., white noise at several different SNRs (with an appropriate quantization of the SNR), as opposed to different noise types multiplied by different SNRs—the large number of possibilities makes data collection extremely challenging in conventional multicondition model training. The advantages of the system, in comparison to electronic noise addition, include the capture of the acoustic coupling between the speech and noise (e.g., the nonlinearities in the handset transducer or the medium), which is assumed to be purely linear in electronic noise addition, and the capture of the effect of the handset transducer on the noise. Additionally, the system may also be able to capture the effect of the distance between the handset and the speech/noise sources, and the effect of room reverberation. A further advance from the system, where applicable, is the replacement of the loudspeaker for speech in Fig. 1 by the true speaker. It is assumed that this will help to further capture the speaker's vocal intensity alternation as a response to ambient noise levels (i.e., the Lombard effect). Other effects, such as the coupling of the transducer to the speech source [40], may also be captured within the system.

The first part of our experiments was concerned with speaker identification. The system shown in Fig. 1 was used to generate the required multicondition training data and the testing data, the latter being obtained by replacing the wide-band noise source with an appropriate test noise source. While capturing the coupling between the speech and environmental noise, the system also captured the reverb characteristics of the recording room. A drawback of the system, as with the other TIMIT-derived databases (e.g., NTIMIT, HTIMIT, CTIMIT), is that it is unable to capture Lombard effects, because the speech material were presented by a loudspeaker, not by a person. Nevertheless, the system is useful as an engineering tradeoff that tries to balance getting more realistic data and getting lots of data. In the second part of our experiments for speaker verification, a realistic noisy speech database was used. The second database

captured realistic noise effects, including the Lombard effect, within the environment it was taken.

As the number of training noise conditions increases, the size of the model increases accordingly based on (1). To limit the size and computational complexity of the model, we can limit the number of mixtures in (1) by pooling the training data from different conditions together and training the model as a usual mixture model to a desired number of mixtures by using the EM algorithm. In this case, the index l in model (1) does not address a specific noise condition any longer, and rather, it is only an index for a mixture component with $P(\Phi_l|S)$ being the mixture weights and $L+1$ being the total number of mixtures for the speaker. This modeling scheme will be examined in our experiments, as a method to reduce the model's complexity through a tradeoff of the model's noise-condition resolution.

III. Speaker Identification Experiments

A. Database and Acoustic Modeling

In the following, we describe our experiments conducted to evaluate the new model for both speaker identification and speaker verification. In the first part of the evaluation, we consider speaker identification. We have developed a new database offering a variety of controlled noise conditions for experiments. This section describes the experiments conducted on this database for closed-set speaker identification. This study is focused on the varieties of noise, and on the development of new methods for generating the training data and reducing the complexity for the new model.

The database contains multicondition training data and test data, both created by using a system illustrated in Fig. 1. To create the multicondition training data for the new model, computer-generated white noise, of the same bandwidth as the speech, was used as the wide-band noise source. Two loudspeakers were used, one playing the wide-band noise and the other playing the clean training utterances. Each training utterance was repeated/recorded in the presence of the wide-band noise $L+1$ times, once without noise (forming Φ_0) and the remaining L times corresponding to L different SNRs (forming $\Phi_1, \ldots, \Phi_L$). In this system, the SNR can be quantified conveniently using the same method as for electronic noise addition. Specifically, for each utterance, the average energy of the clean speech data is calculated, which is used to adjust the average energy of the noise data to be played simultaneously with the speech data subject to a specific SNR. The resulting speech and noise data are then passed to their respective loudspeakers for play and recording, and it is assumed that the recorded noisy speech data can be characterized by the source SNR used to generate the playing data as described above. The test data were generated in exactly the same way as for the training data, by replacing the wide-band noise source in Fig. 1 with a test noise source. As described above, the system captured the acoustic coupling between the speech and noise, which is assumed to be purely additive in electronic noise addition.

The TIMIT database was used as the speech material. This database was chosen primarily for two reasons. First, it was originally recorded under nearly ideal acoustic conditions without noise; this makes it suitable for being used as pristine speech data in our controlled simulation of noisy speech data with the system in Fig. 1. Second, many previous studies on this database, assuming no noise corruption, have shown good recognition accuracy (see, for example, [31], [41], and [44]); this makes it suitable for being used to isolate and quantify the effect of noise on speaker recognition. One disadvantage of the TIMIT database is the lack of handset variability. To make the database also suitable for studying the handset effect, we may follow the way of collecting HTIMIT [40] and use multiple microphones with different characteristics to collect the data in the system of Fig. 1. However, in this study, we focus on the problem of noise effects and assume the use of a single microphone to record the training and test data. In Section IV, we will consider the handset/session variability for speaker verification on a realistic handheld-device database. It is worthwhile to mention that both the PUM approach and the new model described in the paper have been tested previously positively on the SPIDRE database (a subset of the Switchboard corpus) [33], [39]. These early preliminary results were not used in this paper for two reasons: SPIDRE is smaller than TIMIT, and the noise was added artificially while this paper is focused on more realistic noise addition.

The data were recorded in the middle of an office room, with the use an Electret LEM EMU 4535 microphone, placed about 10 cm from the center of the two loudspeakers (i.e., the speech and noise sources) 20 cm away from each other. The room has a dimension of about $4\text{ m} \times 3\text{ m} \times 2.5\text{ m}$ (length, width, and height), with brick walls, a synthetic carpeted floor, and a plaster ceiling. The room is furnished with three computer desks against three walls, plus one bookshelf beside one of the desks. The multicondition training utterances for the new model were recorded in the presence of the wide-band noise at six different SNRs from 10 to 20 dB (increasing 2 dB every step), plus one recording without noise (i.e., clean). While capturing the background noise, the recording system also captured the reverb characteristics of the room. However, reverb effects were not the focus of the paper. Since both the training and testing data were recorded in the same room, we assumed that in our experimental system it is the environmental noise rather than the room reverberation that mainly contributed to the performance degradation.

Six different types of real-world noise data were used, respectively, as the test noise source. These were: 1) jet engine noise; 2) restaurant noise; 3) street noise; 4) polyphonic mobile-phone ring, 5) a pop song with mixed music and voice of a female singer; and 6) a broadcast news segment containing an interview conversation between two male speakers recorded on a highway flyover. Examples of the spectra of these noises are shown in Fig. 2. As can be seen, most of the noises were nonstationary and broad banded, with significant high-frequency components to be accounted for. The durations of these noise files range from about 1 min to about 5 min. For each noise type, we simulated the noisy background by playing the noise in an endless loop, and then obtained the noisy test data by playing and recording the test utterances in the presence of the noise. Data at three different SNRs were recorded: 20, 15, and 10 dB, plus one recording without noise. Because the speech utterances

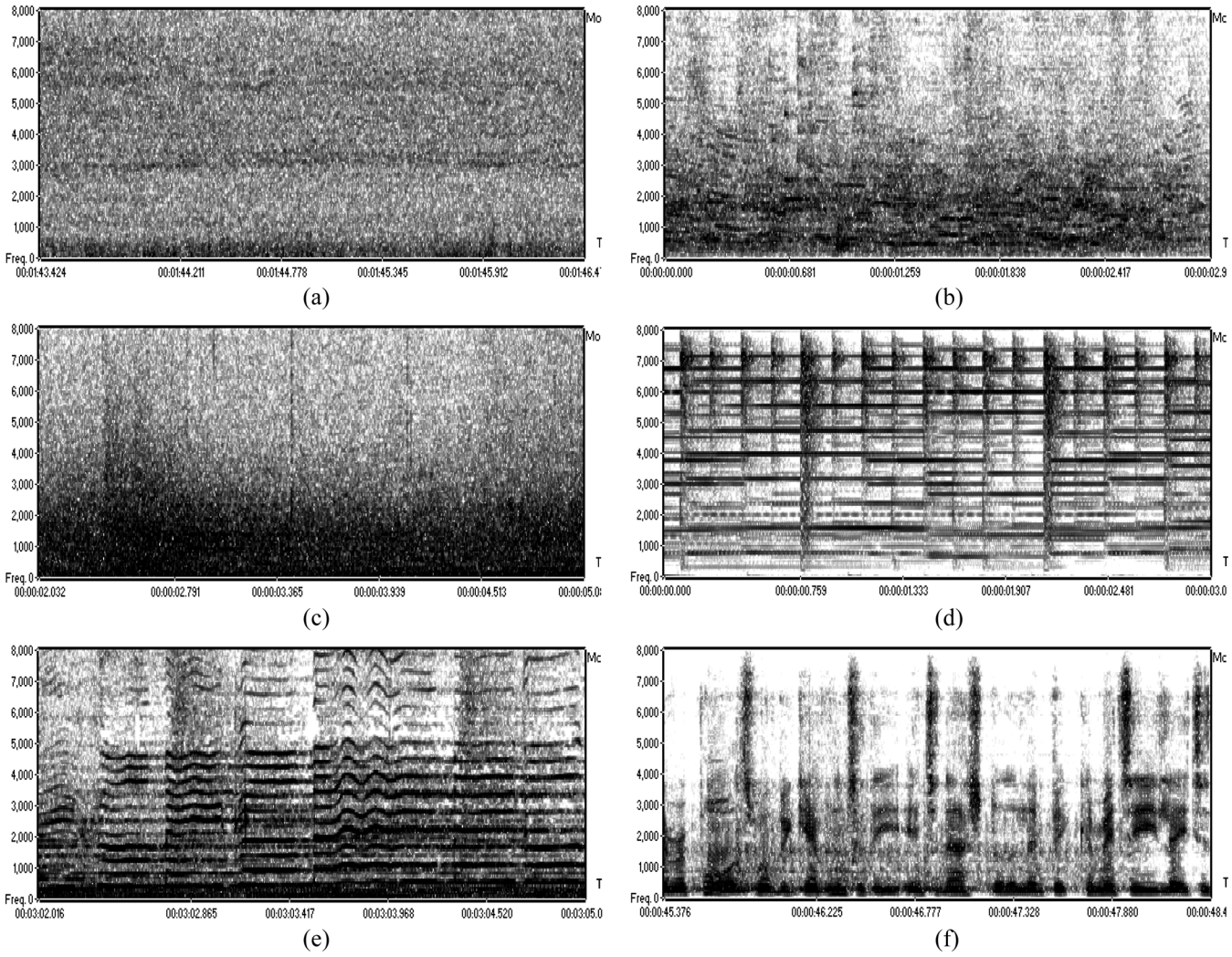


Fig. 2. Noises used in identification experiments, showing the spectra over a period of about three seconds. (a) Jet engine, (b) Restaurant. (c) Street. (d) Mobile-phone ring. (e) Pop song. (f) Broadcast news.

were much shorter than the noise files, each noisy test utterance effectively contained a different portion of the noise file.

The TIMIT database contains 630 speakers (438 male, 192 female), each speaker contributing ten utterances and each utterance having an average duration of about 3 s. Following the practice in [41], for each speaker, eight utterances were used for training, and the remaining two utterances were used for testing. This gives a total of 1260 test utterances across all the 630 speakers. The multicondition training set for each speaker contained 56 utterances (seven SNRs $\times$ eight utterances/SNR). Instead of estimating a separate model for each training SNR condition [which is the model implied in (1)], we pooled all 56 training utterances together and estimated a Gaussian mixture model (GMM) for each speaker, by treating (1) as a normal GMM. As described in Section II-B, by controlling the number of mixtures in this GMM, we gain a control over the the model's complexity. This offers the flexibility to balance noise-condition resolution and computational time.

The speech was sampled at 16 kHz and was divided into frames of 20 ms at a frame period of 10 ms. Each frame was modeled by a feature vector consisting of subband features derived from the decorrelated log filter-bank amplitudes [45], [46]. Specifically, for each frame, a 21-channel Mel-scale filter bank was used to obtain 21 log filter-bank amplitudes, denoted by $(a_1, a_2, \ldots, a_{20}, a_{21})$. These were decorrelated by applying a high-pass filter $H(z) = 1 - z^{-1}$ over a_n, obtaining 20 decorrelated log filter-bank amplitudes, denoted by $(d_1, d_2, \ldots, d_{20}) = (a_2 - a_1, a_3 - a_2, \ldots, a_{21} - a_{20})$. These 20 decorrelated amplitudes were then uniformly grouped into ten subbands, i.e., $(\{d_1, d_2\}, \{d_3, d_4\}, \ldots, \{d_{19}, d_{20}\}) \rightarrow (x_1, x_2, \ldots, x_{10})$, each subband x_n containing two decorrelated amplitudes corresponding to two consecutive filter-bank channels. These ten subbands, with the addition of their corresponding first-order delta components, form a 20-component vector $X = (x_1, x_2, \ldots, x_{10}, \Delta x_1, \Delta x_2, \ldots, \Delta x_{10})$, of a size of 40 coefficients, for each frame.[2]

We implemented three systems all based on the same subband feature format.

[2]Note that we independently model the static components and delta components. This allows the model [i.e., (6)] to only select the dynamic components for scoring. This has been found to be useful for reducing the handset/channel effect, which usually affects the static features more adversely than the dynamic features.

TABLE I
IDENTIFICATION ACCURACY (%) FOR THE NEW MODEL AND BASELINE MULTICONDITION MODEL BSLN-MUL TRAINED USING SIMULATED, ACOUSTICALLY MIXED MULTICONDITION DATA AT SEVEN DIFFERENT SNRS, AND FOR THE BASELINE MODEL BSLN-CLN TRAINED USING CLEAN DATA, ALL USING SUBBAND FEATURES. THE LAST CATEGORY SHOWS THE ACCURACY BY A BASELINE GMM USING FULL-BAND MFCC, TRAINED ON THE MULTICONDITION DATA (MUL) AND CLEAN DATA (CLN), RESPECTIVELY. THE NUMBER ASSOCIATED WITH EACH MODEL INDICATES THE NUMBER OF GAUSSIAN MIXTURES IN THE MODEL

Noise	SNR	New model			BSLN-Mul	BSLN-Cln	Fullband MFCC	
	(dB)	32	64	128	128	32	Mul 128	Cln 32
Clean		90.64	94.84	96.51	95.79	98.41	93.81	95.87
Engine	20	83.81	87.06	88.89	86.35	62.46	57.54	22.14
	15	78.26	81.75	81.59	77.62	29.05	55.56	6.91
	10	51.27	52.30	51.35	53.57	7.78	43.49	1.35
Restaurant	20	85.87	91.27	93.89	94.44	93.10	79.13	77.06
	15	80.56	85.95	88.33	87.46	78.97	54.13	41.27
	10	67.54	73.25	75.08	67.70	43.57	22.70	12.30
Street	20	86.75	91.27	92.86	94.29	91.83	71.83	70.08
	15	79.76	85.08	86.51	86.83	70.32	40.24	37.30
	10	61.11	63.57	64.05	68.17	34.60	12.30	10.71
Mobile	20	73.57	80.64	84.68	68.02	56.90	52.70	24.60
phone ring	15	63.65	72.30	76.35	46.90	34.05	40.56	9.84
	10	48.10	57.38	62.46	26.43	15.56	26.27	3.25
Pop	20	87.54	92.22	93.41	86.19	88.57	81.75	83.17
song	15	78.26	85.71	88.07	64.44	66.98	62.46	60.32
	10	58.49	64.21	67.70	33.65	30.87	34.13	31.43
Broadcast	20	87.22	92.54	93.89	82.78	84.92	77.78	77.46
news	15	79.05	86.03	88.97	59.84	61.75	55.16	57.70
	10	57.87	66.75	70.00	27.62	26.19	29.76	29.05

1) BSLN-Cln: A baseline GMM trained on clean data and tested using the full set of subband features, with 32 mixtures per speaker.
2) BSLN-Mul: A baseline GMM trained on the simulated multicondition data and tested using the full set of subband features, with 128 Gaussian mixtures per speaker.
3) New model: The proposed model (6), trained on the simulated multicondition data and tested using optimally selected subband features for each training condition, with 32, 64, and 128 Gaussian mixtures, respectively, per speaker.

Additionally, for comparison, we also implemented a baseline GMM system that used conventional full-band Mel-frequency cepstral coefficients (MFCC) instead of the above subband features. In the system, each frame was modeled by a 24-component vector, consisting of 12 MFCC plus 12 first-order delta MFCC, derived from a 26-channel Mel-scale filter bank (this corresponds to the default configuration used in the HTK system for the TIMIT database).

B. Identification Results

Table I presents the identification accuracy obtained by the various models in all the tested conditions. The accuracy of 98.41% for the clean test data by the clean baseline BSLN-Cln represents one of the best identification results we have ever obtained on the TIMIT database. This may indicate that the distortion on the speech signal imposed by our play/recording procedure for data collection (Fig. 1) is negligible and that the acoustic features and models used to characterize the speakers are adequate.

For the new model, given a noise/SNR condition, the accuracy improved as the number of mixtures increased because of a higher noise-level resolution. We only experienced exceptions for the engine noise in the 10/15-dB SNR cases, which showed a small fluctuation in accuracy when the number of mixtures increased from 64 to 128. With 128 mixtures (on average, about $128/7 \simeq 18$ mixtures per SNR condition), the new model was able to outperform the baseline model BSLN-Cln in all tested noisy conditions, with a small loss of accuracy for the noise-free condition. Compared to the baseline multicondition model BSLN-Mul, the new model obtained improved accuracy in the majority of test conditions. As expected, the improvement is more significant for those noise types that are significantly different from the wide-band white noise used to train the new model and the BSLN-Mul model. In our experiments, for example, these noises include the mobile phone ring, pop song, and broadcast news, all showing very different spectral structures from the white noise spectral structure (Fig. 2). For these noises, the new model improved over BSLN-Mul by focusing less on the mismatched noise characteristics. However, for those noises that are close to wide-band white noise and thus can be well modeled by BSLN-Mul, the new model offered less significant improvement or no improvement. In our experiments, these noises include the engine noise, restaurant

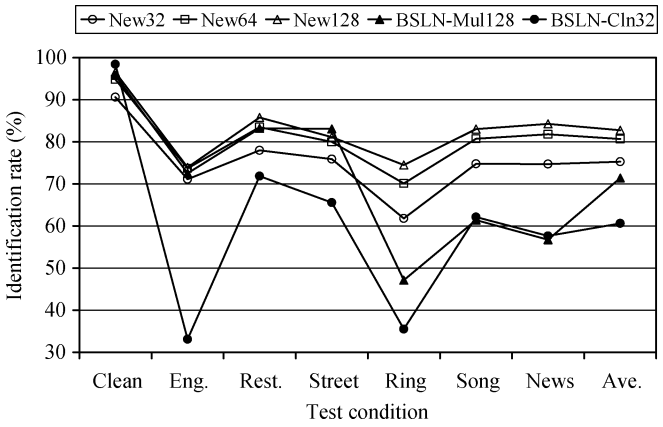


Fig. 3. Identification accuracy in clean and six noisy conditions averaged over SNRs between 10–20 dB, and the overall average accuracy across all the conditions, for the new model and the BSLN-Mul model trained using simulated, acoustically mixed multicondition data at seven different SNRs, and for the BSLN-Cln model trained using clean data. The number associated with each model indicates the number of Gaussian mixtures in the model.

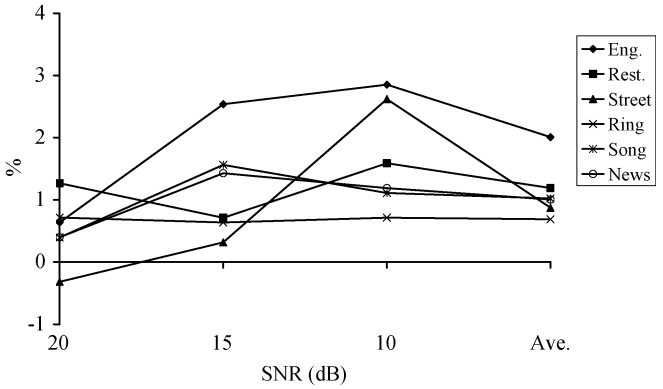


Fig. 4. Absolute improvement in identification accuracy for the new model trained on multicondition data with acoustically added noise, compared to trained on multicondition data with electronically added noise, tested on data with acoustically added noise, with 128 Gaussian mixtures per speaker.

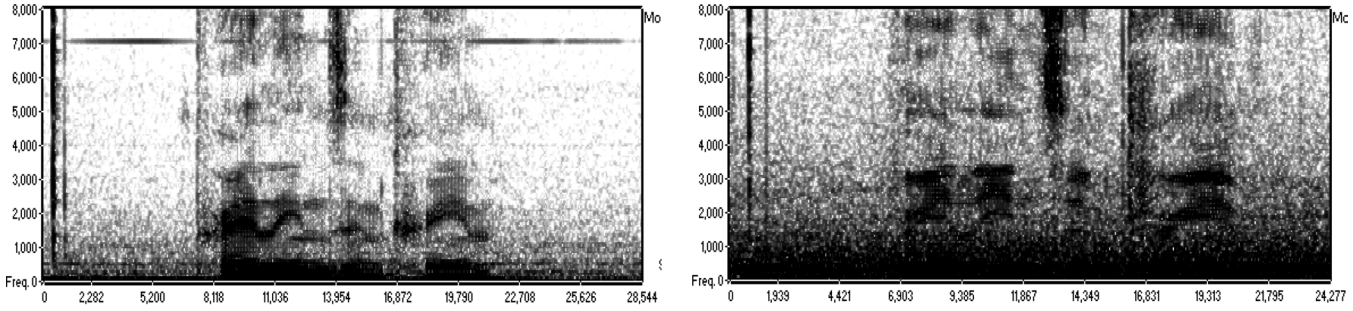
noise, and street noise.[3] For these noises, the new model and the BSLN-Mul model achieved similar performances, and, because of being trained in the well-matched wide-band noise, BSLN-Mul performed significantly better than BSLN-Cln trained only using clean data. The improvement of BSLN-Mul over BSLN-Cln was much less significant for the other three mismatched noises—mobile phone ring, pop song, and broadcast news. Fig. 3 shows the average performance by the three systems across all the tested clean/noisy conditions. All the three new models, with 32, 64, and 128 mixtures, respectively, showed better average performance than the other two systems, indicating the potential of the new system for dealing with a wider variety of noisy conditions. The relative processing time for the BSLN-Mul model with 128 mixtures compared to the new model also with 128 mixtures was about 1:6. This ratio dropped almost linearly to about 1:3 for the new model with 64 mixtures and to about 1:1.5 for the new model with 32 mixtures. The last category of Table I shows the identification accuracy obtained by the baseline GMM using full-band MFCC. It is noticed that on this database, the full-band, MFCC-based baseline (Mul, Cln) performed poorer than the corresponding subband-based baseline (BSLN-Mul, BSLN-Cln) in the majority of test conditions. We also tested the application of sentence-level cepstral mean normalization to the full-band MFCC and found no improvement in identification accuracy.

C. Acoustic Noise Addition Versus Electronic Noise Addition

In the above experiments, the multicondition training data for the new model were created using the system shown in Fig. 1, in which the wide-band noise was acoustically mixed into the clean training data; the noisy test data were also created in the same way, i.e., acoustic noise addition (ANA). This model is different from the commonly used additive-noise model, which assumes, among other assumptions, that the coupling of speech and background noise is a linear sum of the clean speech signal and the noise signal. The additive-noise model allows the simulation of noisy speech by electronically adding noise to clean speech, i.e., electronic noise addition (ENA). In the following we describe an experiment to compare ENA and ANA for being used to generate the multicondition training data for the new model. Specifically, in the experiment, we assumed that the test data were generated in the same way as above using ANA, but the multicondition training data were generated using ANA and ENA, respectively. This comparison is of interest because it could offer an idea about how accurate the additive-noise model is for characterizing acoustically coupled noisy speech signals, in terms of the recognition performance. To keep the other conditions exactly the same in the comparison, the noise data associated with each training utterance in ANA were saved and later played/recorded alone without presence of speech; the recorded pure noise was then added electronically to the previously recorded clean speech to form a noisy training utterance. This procedure minimized the SNR difference between the data generated by the two methods and introduced the same transducer and room reverb effects on the resulting noisy training data.

Fig. 4 shows the absolute improvement in identification accuracy obtained by ANA-based training over ENA-based training, for the noisy test signals generated with an ANA model. Small, positive improvements were observed in all tested conditions except for the 20-dB street noise case. The results in Fig. 4 indicate little degradation from ANA to ENA, appearing to suggest that given the speech and noise signals, ENA is a reasonably accurate model for their physical coupling. Research should thus focus on the factors that directly modify the signal sources (e.g., Lombard effects [47], [48]), and the factors that alter the characteristics of the observed signals (e.g., handset/channel effects [40], room reverberation [49], etc.). Later in Section V we will discuss a possible extension of the new model and the training data collection system for modeling new forms of signal distortion.

[3]We have conducted an extra experiment that is not included in the paper. In the experiment, we trained a baseline multicondition model by replacing the wide-band noise in Fig. 1 with each of the three test noises—engine, restaurant, and street—at 20, 15, and 10 dB, and thereby created a model that almost exactly matches the test conditions with the three noises. The identification accuracy produced by this "matching" model for the matched noise conditions is very similar to the accuracy obtained by the BSLN-Mul model. This indicates the similarity in characteristics between the three noises and the simulated wide-band noise.



(a) (b)

Fig. 5. Spectra of utterances recorded in (a) office and (b) street intersection, using the internal microphone.

IV. Speaker Verification Experiments

A. Database and Acoustic Modeling

This section describes further experiments to evaluate the new model with the use of real-world application data. The MIT Mobile Device Speaker Verification Corpus [50] was used in the experiments (which extend previous results reported in [51]). The database was designed for speaker verification with limited enrollment data, and was collected using a handheld-device in realistic conditions with the use of an internal microphone and an external headset. The database contains 48 enrolled speakers (26 male, 22 female) and 40 impostors (23 male, 17 female), each reciting a list of name and ice-cream flavor phrases. The part of the database containing the ice-cream flavor phrases was used in the experiments. There were six phrases rotated among the enrolled speakers, with each speaker reciting an assigned phrase four times for training and four times for verification. The training and test data were recorded in separate sessions, involving the same or different background/microphone conditions and different phrase rotation. The same practice applies to the impostors, with each impostor repeating an assigned phrase four times in each given background/microphone condition with condition-varying phrase rotation. The impostors saying the same phrase as an enrolled speaker were grouped to form the impostor trials for that enrolled speaker. Then, in each test, there were a total of 192 enrolled speaker trials and a slightly varying number of impostor trials ranging from 716 to 876 depending on the test conditions.

We considered the data collected in two different environments: office (with a low level of background noise) and street intersection (with a higher level of background noise). Fig. 5 shows the typical characteristics of the environments. We assumed that the speaker models were trained based on the office data and tested in matched and mismatched conditions without assuming prior information about the test environments. The office data served as Φ_0, from which multicondition training sets $\Phi_1, \ldots, \Phi_L$ were generated by introducing different corruptions into Φ_0. In our experiments, we tested the addition of wide-band noise and narrow-band noise, respectively, to the clean training data for creating the noisy training data sets. The noise was added electronically. The wide-band noise was obtained by passing a white noise through a low-pass filter with the same bandwidth as the speech spectrum, and the narrow-band noise was obtained in the same way but with a lower 3-dB cutoff frequency, i.e., 800 Hz, for the low-pass filter. The latter simulates the weakening high-frequency components for the noise, as may be seen in Fig. 5. We have tested other cutoff frequencies within the range 700–2000 Hz for the narrow-band training noise and found that they offered similar performances. In the following, we first present the experimental results for the separate use of the wide-band noise and the narrow-band noise for training the models. It was found that wide-band training noise was not the best choice for this database with relatively weak high-frequency noise components. However, we have seen earlier in Section III that wide-band training noise is needed for dealing with noise sources with significant high-frequency components. In the final part of this experiment, we demonstrate a model built upon mixed wide-band and narrow-band noise training, to optimize the performance for varying noise bandwidths.

We added the simulated noise to each training utterance at nine different SNRs between 4–20 dB (increasing 2 dB every step). This gives a total of ten training conditions (including the no corruption condition), each characterized by a specific SNR. We treated the problem as text-dependent speaker verification, and modeled each enrolled speaker using an eight-state HMM, with each state in each condition (i.e., $p(X|S, \Phi_l)$, which now models the observation likelihood in state S within a speaker's HMM) being modeled by two diagonal-Gaussian mixtures. Additionally, three states with 16 mixtures per state were used to account for the beginning and ending backgrounds within each utterance; these states were tied across all the speakers. The $p(X|S, \Phi_l)$ for different Φ_l were combined based on (1) assuming a uniform prior $P(\Phi_l|S)$; no model size reduction was considered in this case because of the small number of mixtures in each $p(X|S, \Phi_l)$. The signals were sampled at 16 kHz and were modeled using the same frame/subband feature structure as described in Section III-A, with an additional sentence-level mean removal for the subband features (similar to cepstral mean subtraction).

We implemented three systems all based on the same subband feature format, and all having the same state-mixture topology as described above.

1) BSLN-Cln: A baseline system trained on "clean" (office) data.
2) BSLN-Mul: A baseline system trained on the simulated multicondition data.

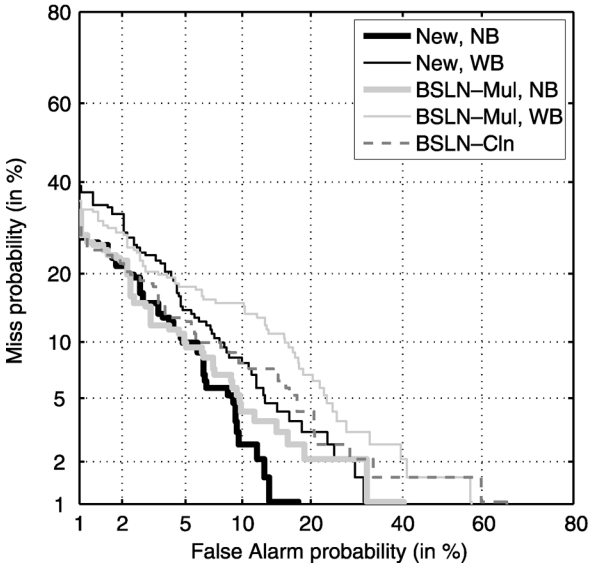


Fig. 6. DET curves in matched training and testing: Office/headset, for the new model and the BSLN-Mul model trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for the BSLN-Cln model trained using clean data.

3) New model: The proposed model (6) trained on the simulated multicondition data.

Two cases were further considered for the new model and the BSLN-Mul model: 1) the use of wide-band noise and 2) the use of narrow-band noise to generate the multicondition training data.

B. Verification Results

We first compared the three systems assuming matched condition training and testing, both in the office environments with the use of a headset. Fig. 6 presents the detection-error-tradeoff (DET) curves, for the new model and the BSLN-Mul model trained using narrow-band noise (NB) and wide-band noise (WB), respectively, and for the BSLN-Cln model trained using clean data. The office data are not perfectly clean, often with burst noise at the time the microphone being switched on/off and some random background noise. Fig. 6 indicates the usefulness for reducing the mismatch by training the models in narrow-band noise, as seen for the better performances obtained by the two multiconditionally trained, narrow-band noise-based models New (NB) and BSLN-Mul (NB), over the single-conditionally trained model BSLN-Cln. However, training the models using the wide-band noise hurt the performance, particularly for BSLN-Mul (WB), due to the serious mismatch between the training and testing conditions. The new model improved the situation by ignoring some of the mismatched data, and offered better performance over its counterpart BSLN-Mul in both narrow-band noise and wide-band noise training conditions. Table II summarizes the equal error rates (EERs) associated with each system in different training/testing conditions. As shown in the table, for this matched condition training/testing case (index: OH-OH), the new model obtained lower EERs than the other systems assuming the same information about the test condition.

TABLE II
EQUAL ERROR RATES (%) FOR THE NEW MODEL AND THE BSLN-MUL MODEL TRAINED USING SIMULATED NARROW-BAND NOISE (NB), WIDE-BAND NOISE (WB) AND COMBINATION (NB+WB) AT TEN DIFFERENT SNRS, AND FOR THE BSLN-CLN MODEL TRAINED USING CLEAN DATA (INDEX: O—OFFICE, S—STREET INTERSECTION, H—HEADSET, I—INTERNAL MICROPHONE)

Training-Testing condition	New model			BSLN-Mul		BSLN-Cln
	NB	WB	NB+WB	NB	WB	
OH - OH	6.50	8.45	7.79	7.29	12.65	8.85
OI - SI	11.98	15.63	13.51	15.63	23.96	20.83
OI - SH	14.06	17.71	14.62	22.40	30.73	30.21

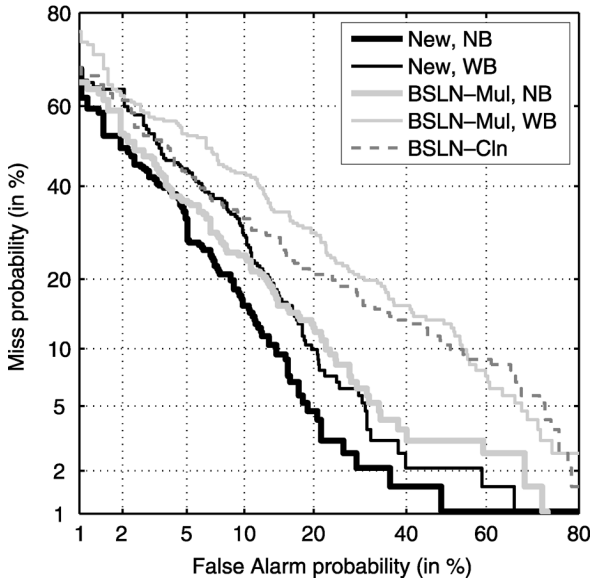


Fig. 7. DET curves with mismatch in environments: training—office, testing—street intersection, both using internal microphone, for the new model and the BSLN-Mul model trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for the BSLN-Cln model trained using clean data.

Next, we tested the three systems assuming there is training/testing mismatch in environments but no mismatch in microphone type. The models were trained using the office data and tested using the street intersection data, both collected using the internal microphone. Fig. 7 shows the DET curves, and Table II shows the corresponding EERs (index: OI-SI). The new model offered improved performance, reducing the EER by 42.5/24.9% (NB/WB) as compared to BSLN-Cln. While the narrow-band noise-based BSLN-Mul (NB) improved over BSLN-Cln, the wide-band noise-based BSLN-Mul (WB) performed worse than BSLN-Cln, with a higher EER. This is due to the severe mismatch in the noise characteristics (e.g., bandwidth) between the training and testing. This mismatch was reduced in the new model by focusing on the matching subbands. As seen, the new model (WB) trained on the less matched wide-band noise performed similarly to BSLN-Mul (NB) trained on the better matched narrow-band noise, in terms of the EER. The new model (NB/WB) reduced the EER by 23.4/34.8% as compared to the corresponding BSLN-Mul (NB/WB).

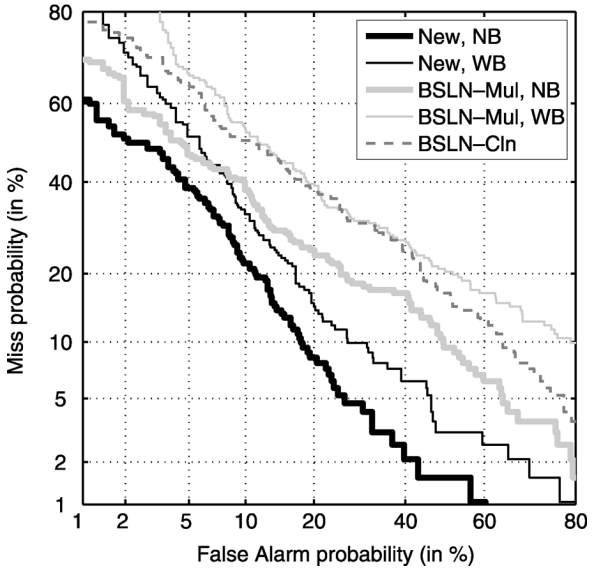


Fig. 8. DET curves with mismatch in both environments and microphones: training—office/internal microphone, testing—street intersection/headset, for the new model and the BSLN-Mul model trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for the BSLN-Cln model trained using clean data.

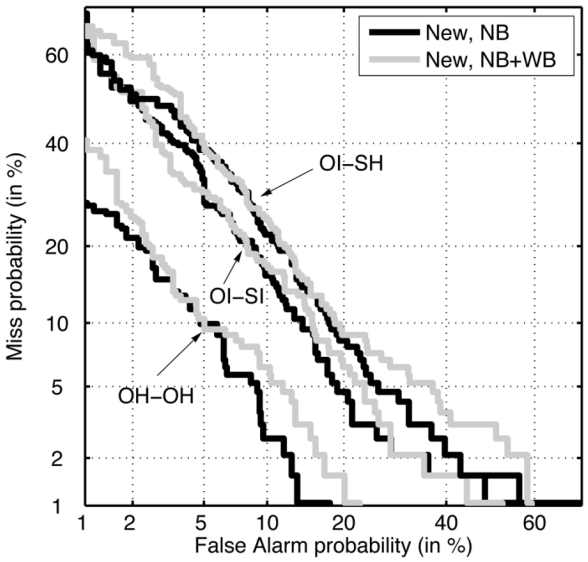


Fig. 9. Comparison between the new models trained using simulated narrow-band noise (NB) and mixed narrow-band noise and wide-band noise (NB+WB), for different training-testing environment/microphone conditions (Index: O—office, S—street intersection, H—headset, I—internal microphone).

Further experiments were conducted assuming mismatch in both environments and microphone types. The models were trained using the office data with an internal microphone and tested using the street intersection data with a headset. Fig. 8 presents the DET curves with the corresponding EERs shown in Table II (index: OI-SH). Again, the new model offered improved performance over both BSLN-Cln and BSLN-Mul. Compared to BSLN-Cln, the new model (NB/WB) reduced the EER by 53.4/41.4%, and compared to BSLN-Mul (NB/WB), the reductions were 37.2/42.4%. It is noted that in this case of combined mismatch, the new model (WB) offered lower EER than BSLN-Mul (NB)—the latter was trained using narrow-band noise that better matched the test environment than the wide-band noise (WB). Therefore, the new model resulted in the lowest EERs among all the tested systems.

The above experimental results reveal that a knowledge of the noise bandwidth could help improve the new model's performance. By training the model using low-pass filtered white noise matching the noise bandwidth, the model would ideally pick up information both from the noisy subbands (due to the compensation) and from the remaining little corrupted subbands (through matched clean subbands between the model and data), and therefore obtain more information, i.e., a larger subset X_l in (2), for recognition. Otherwise, if the model $p(X|S,\Phi_l)$ is trained using wide-band white noise, the information from the clean subbands of the test signal would have to be ignored to reduce the model-data mismatch, resulting in a loss of information. Without assuming knowledge of the noise bandwidth, we may consider building the model by using mixed noise data, with increasing bandwidths, to offer improved accuracy for modeling band-limited noise while providing coverage for wide-band noise. In the following, we show an example by combining the two new models described above, one trained on the narrow-band noisy data and the other on the wide-band noisy data, to form a single model based on (1). The results are shown in Fig. 9, for all the above examined training/testing conditions and including a comparison with the narrow-band noise-based model (NB). As can be seen, the combined model improved over the wide-band noise-based model (WB), performed similarly to the narrow-band noise-based model (NB), and, at the same time, retained the potential of the wide-band noise-based model (WB) for dealing with wide-band noise corruption. The EERs for the combined model are included in Table II.

As mentioned earlier, multicondition model training using added noise at various SNRs to account for unknown noise sources has been studied previously in speech recognition (e.g., [37]). The above experimental results indicate that, compared to clean-data training, multicondition training may or may *not* offer improved performance, depending on how well the training noise data match the testing noise data in characteristics. The training/testing mismatch can be reduced, and hence improved robustness obtained, by combining multicondition training with a missing-feature model, as evident by the performance differences between the new model and the BSLN-Mul model.

V. Summary

This paper investigated the problem of speaker recognition in noisy conditions assuming absence of information about the noise. We described a method that combines multicondition model training and missing-feature theory to model noise with unknown temporal-spectral characteristics. Multicondition training is conducted using simulated noisy data of simple noise characteristics, providing a coarse compensation for the noise, and missing-feature theory is applied to refine the compensation by ignoring noise variation outside the given training conditions, thereby accommodating training and testing mismatch.

We studied the new model for both speaker identification and speaker verification. The research is focused on new methods for creating multicondition training data to model noisy speech, on the combination of training data of different characteristics to optimize the recognition performance, and on the reduction of the model's complexity by training the model as a usual GMM. So far we have experimented the addition of wide-band white noise, and a combination of wide-band white noise and low-pass filtered white noise, to cover various noises of different spectral shapes and bandwidths. We expect further improved simulation accuracy by additionally including realistic noises into the corruption, depending on the expected environments. Two databases were used to evaluate the new algorithm. The first was a noisy TIMIT database obtained by rerecording the data in various controlled noise conditions, used for an experimental development of the new model with a focus on the varieties of noise. The second was a handheld-device database collected in realistic noisy conditions, used to further validate the model by testing on a real-world application. Experiments on both databases have shown improved noise robustness for the new model, in comparison to baseline systems trained on the same amount of information. An additional experiment was conducted to compare the traditional additive-noise model and acoustic noise addition for modeling noisy speech. Acoustic noise addition is feasible in the new model due to its potential of modeling arbitrary noise conditions with the use of a limited number of simulated noise conditions. Currently, we are considering an extension of the principle of the new model to model new forms of signal distortion, e.g., handset variability, room reverberation, and distant/moving speaking. We will modify the system in Fig. 1 so that it can be used to collect training data for these factors. To make the task tractable, these factors can be "quantized" as we did for the noise bandwidth and SNR. Missing-feature approaches will be used to deemphasize the mismatches while exploiting the matches arising from the quantized data.

REFERENCES

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[3] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.

[4] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[5] S. van Vuuren, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch," in *Proc. ICSLP'96*, Philadelpia, PA, 1996, pp. 1788–1791.

[6] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe, "Global optmization of a neural network-hidden markov model hybrid," *IEEE Trans. Neural Netw.*, vol. 3, no. 2, pp. 252–259, Mar. 1992.

[7] S. Euler, "Integrated optimization of feature transformation for speech recognition," in *Proc. Eurospeech'95*, Madrid, Spain, 1995, pp. 109–112.

[8] M. Rahim, Y. Bengio, and Y. Lecun, "Discriminative feature and model design for automatic speech recognition," in *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 75–78.

[9] L. P. Heck, Y. Konig, M. K. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Commun.*, vol. 31, pp. 181–192, 2000.

[10] R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition—A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[11] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Magnitude-only estimation of handset nonlinearity with application to speaker recopgnition," in *Proc. ICASSP'98*, Seattle, WA, 1998, pp. 745–748.

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey—The Speaker Recognition Workshop*, Crete, Greece, 2001, pp. 213–218.

[13] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. ICASSP'02*, Orlando, FL, 2002, pp. 681–684.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[15] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP'03*, Hong Kong, China, 2003, pp. 49–52.

[16] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, pp. 42–54, 2000.

[17] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 554–568, Sep. 1999.

[18] R. Teunen, B. Shahshahani, and L. P. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. ICSLP'00*, Beijing, China, 2000, pp. 495–498.

[19] L. F. Lamel and J. L. Gauvain, "Speaker verification over the telephone," *Speech Commun.*, vol. 31, pp. 141–154, 2000.

[20] K. K. Yiu, M. W. Mak, and S. Y. Kung, "Environment adaptation for robust speaker verification," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003, pp. 2973–2976.

[21] J. Ortega-Garcia and L. Gonzalez-Rodriguez, "Overview of speaker enhancement techniques for automatic speaker recognition," in *Proc. ICSLP'96*, Philadelpia, PA, 1996, pp. 929–932.

[22] Suhadi, S. Stan, T. Fingscheidt, and C. Beaugeant, "An evaluation of VTS and IMM for speaker verification in noise," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003, pp. 1669–1672.

[23] M. J. F. Gales and S. Young, "HMM recognition in noise using parallel model combination," in *Proc. Eurospeech'93*, Berlin, Germany, 1993, pp. 837–840.

[24] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Comput. Speech Lang.*, vol. 10, pp. 107–116, 1996.

[25] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. ICASSP'01*, Salt Lake City, UT, 2003, pp. 457–460.

[26] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. ICASSP'97*, Munich, Germany, 1997, pp. 835–838.

[27] C. Cerisara, L. Rigaziob, and J.-C. Junqua, "a-Jacobian environmental adaptation," *Speech Commun.*, vol. 42, pp. 25–41, 2004.

[28] L. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust speaker recognition through acoustic array processing and spectral normalization," in *Proc. ICASSP'97*, Munich, Germany, 1997, pp. 1103–1106.

[29] I. McCowan, J. Pelecanos, and S. Scridha, "Robust speaker recognition using microphone arrays," in *Proc. A Speaker Odyssey—The Speaker Recognition Workshop*, Crete, Greece, 2001, pp. 101–106.

[30] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," in *Proc. ICASSP'98*, Seattle, WA, 1998, pp. 121–124.

[31] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modelling," *Speech Commun.*, vol. 31, pp. 89–106, 2000.

[32] J. Ming, "Universal compensation—An approach to noisy speech recognition assuming no knowledge of noise," in *Proc. ICASSP'04*, Montreal, QC, Canada, 2004, pp. I.961–I.964.

[33] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions—A universal compensation approach," in *Proc. ICASSP'05*, Philadelphia, PA, 2005, pp. 617–620.

[34] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP'96*, Philadelpia, PA, 1996, pp. 426–429.

[35] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on partially corrupted speech," in *Proc. ICSLP'96*, Philadelpia, PA, 1996, pp. 462–465.

[36] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP'87*, Dallas, TX, 1987, pp. 705–708.
[37] L. Deng, A. Acero, M. Plumpe, and X.-D. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP'00*, Beijing, China, 2000, pp. 806–809.
[38] J. Ming, P. Jancovic, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 403–414, Sep. 2002.
[39] J. Ming, J. Lin, and F. J. Smith, "A posterior union model with applications to robust speech and speaker recognition," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–12, 2006, Article ID 75390.
[40] D. A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects," in *Proc. ICASSP'97*, Munich, Germany, 1997, pp. 1535–1538.
[41] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.
[42] K. L. Brown and E. B. George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," in *Proc. ICASSP'95*, Detroit, MI, 1995, pp. 105–108.
[43] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," in *Proc. ICASSP'90*, Albuquerque, NM, 1990, pp. 109–112.
[44] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Commun.*, vol. 24, pp. 193–209, 1998.
[45] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of the filter-bank energies in speech recognition," in *Proc. Eurospeech'95*, Madrid, Spain, 1995, pp. 1381–1384.
[46] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proc. Eurospeech'99*, Budapest, Hungary, 1999, pp. 85–88.
[47] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizer," *J. Acoust. Soc. Amer.*, vol. 93, pp. 510–524, 1993.
[48] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, pp. 151–173, 1996.
[49] D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation," in *Proc. ICSLP'96*, Trento, Italy, 1996, pp. 1329–1332.
[50] R. Woo, A. Park, and T. J. Hazen, "The MIT mobile device speaker verification corpus: Data collection and preliminary experiments," in *Proc. IEEE Odyssey 2006—The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006, pp. 1–6 [Online]. Available: http://groups.csail.mit.edu/sls/mdsvc
[51] J. Ming, T. J. Hazen, and J. R. Glass, "Speaker verification over handheld devices with realistic noisy speech data," in *Proc. ICASSP'06*, Toulouse, France, 2006, pp. 637–640.

Ji Ming (M'97) received the B.Sc. degree from Sichuan University, Chengdu, China, in 1982, the M.Phil. degree from Changsha Institute of Technology, Changsha, China, in 1985, and the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 1988, all in electronic engineering.

He was Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. Since 1993, he has been with the Queen's University Belfast, Belfast, U.K., where he is currently a Professor in the School of Electronics, Electrical Engineering, and Computer Science. From 2005 to 2006, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. His research interests include speech and language processing, image processing, and pattern recognition.

Timothy J. Hazen (M'04) received the S.B., S.M., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1991, 1993, and 1998, respectively.

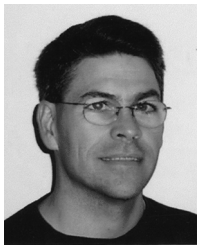
From 1998 to 2007, he was a Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory. He is currently serving as a member of the technical staff at MIT Lincoln Laboratory. His research interests include automatic speech recognition, automatic person identification, multimodal speech processing, and conversational speech systems.

Dr. Hazen has also served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2004 to 2007.

James R. Glass (M'78–SM'06) received the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1985, and 1988, respectively.

After starting in the Speech Communication Group at the MIT Research Laboratory of Electronics, he has worked at the Laboratory for Computer Science, now the Computer Science and Artificial Intelligence Laboratory (CSAIL), since 1989. Currently, he is a Principal Research Scientist at CSAIL, where he heads the Spoken Language Systems Group. He is also a Lecturer in the Harvard-MIT Division of Health Sciences and Technology. His primary research interests are in the area of speech communication and human–computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, supervised students, and published extensively in these areas.

He has previously been a member of the IEEE Signal Processing Society Speech Technical Committee and an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.



Douglas A. Reynolds (M'86–SM'98) received the B.E.E. degree (with highest honors) and the Ph.D. degree in electrical engineering both from the Georgia Institute of Technology, Atlanta.

He joined the Speech Systems Technology Group (now the Information Systems Technology Group), Massachusetts Institute of Technology Lincoln Laboratory in 1992. Currently, he is a Senior Member of Technical Staff, and his research interests include robust speaker and language identification and verification, speech recognition, and general problems in signal classification and clustering.

Dr. Reynolds is a senior member of the IEEE Signal Processing Society and a cofounder and member of the steering committee of the Odyssey Speaker Recognition Workshop.