Robust Speaker Recognition in Unknown Noisy Conditions

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

EDICS: SPE-SPKR

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 of the noise. This paper describes a method, named *universal compensation* (UC), that combines multi-condition training and the missing-feature method to model noises with unknown temporal-spectral characteristics. Multi-condition training is conducted using simulated noisy data with limited noise varieties, providing a "coarse" compensation for the noise, and the missing-feature method refines the compensation by ignoring noise variations 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 UC model for real-world applications. These include the generation of multi-condition training data to model real-world noisy speech, the combination of different training data to optimize the recognition performance, and the reduction of the model's complexity. Two databases were used to test the UC algorithm. The first is a re-development of the TIMIT database by re-recording the data in the presence of various noises, used to test the model for speaker identification with a focus on the noise varieties. The second is a handheld-device database collected in realistic noisy conditions, used to further validate the model on the real-world data for speaker verification. The new model was compared to baseline systems and has shown improved identification and verification performance.

J. Ming is with the School of Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: j.ming@qub.ac.uk, phone: 44-28-90974723; fax: 44-28-90975666).

T. J. Hazen and J. R. Glass are with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, U.S.A. (e-mail: hazen/glass@csail.mit.edu; phone: 1-617-253-4672/1640; fax: 1-617-258-8642).

D. A. Reynolds is with the MIT Lincoln Laboratory, Lexington, MA 02420, U.S.A. (email: dar@ll.mit.edu; phone: 1-781-981-4494; fax: 1-781-981-0186).

The work was sponsored in part by Intel Corporation, and in part by 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.

I. INTRODUCTION

Accurate speaker recognition is made 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 towards reducing the effect of handset/channel mismatch. Linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains. Examples of the feature compensation methods include well-known filtering techniques such as cepstral mean subtraction or RASTA (e.g. [1]–[3]), discriminative feature design with neural networks [4] and various feature transformation methods such as nonlinear spectral magnitude normalization, feature warping and short-time Gaussianization (e.g. [5]–[7]). Score-domain compensation aims to remove handset-dependent biases from the likelihood ratio scores. The most prevalent methods include H-Norm [8], Z-norm [9] and T-Norm [10]. Examples of the model-domain compensation methods include the speaker-independent variance transformation [11], and the transformation for synthesizing supplementary speaker models for other channel types from multi-channel training data [12]. Additionally, channel mismatch has also been dealt with by using model adaptation methods, which effectively use new data to learn channel characteristics (e.g. [13], [14]).

To date, research has targeted the impact of environmental noise through filtering techniques such as spectral subtraction or Kalman filtering [16], [17], assuming *a priori* knowledge of the noise spectrum. Other techniques rely on a statistical model of the noise, for example, PMC (parallel model combination) [18], [19], or on the use of microphone arrays [20], [21]. Recent studies on the missing-feature method 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. [22], [23]). 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 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. In this paper, a method, namely *universal compensation* (UC), is proposed. The UC technique is an extension of the missing-feature method, 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. The UC technique involves a combination of the multi-condition training method and the missing-feature method. Multi-condition training, with simulated noisy data of limited noise varieties, serves as the first step to provide a "coarse" compensation for the noise. The missing-feature method serves as the second step to fine "tune" the compensation by ignoring noise variations outside given training conditions, thereby accommodating mismatches between the simulated training noise condition and the realistic test noise condition. The UC technique represents an effort to model arbitrary noise conditions by using a limited number of simulated noise conditions.

As preliminary studies, the UC method was first tested for speech recognition (e.g. [24]) and later for speaker identification [25], both using artificially synthesized noisy speech data. This paper extends the previous research by focusing on two problems: 1) improving the model's capability for modeling realistic noisy speech, and 2) exploring the application of the model towards real-world problems for both speaker identification and speaker verification. More specifically, we will study new methods for generating multi-condition training data for the UC model 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. Two databases are used to evaluate the proposed model. The first is a re-development of the TIMIT database by re-recording the data in various controlled noise conditions, with a focus on the noise varieties. The UC model, along with the proposed methods for generating the training data and reducing the model complexity, was tested and developed on this database for speaker identification. The second is a realistic handheld-device database collected in realistic noisy conditions. The UC 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 test on a real-world application.

The remainder of this paper is organized as follows. Section 2 describes the UC method and the methods for generating the training data and controlling the model's complexity. Section 3 presents the experimental results for speaker identification on the noisy TIMIT database, and Section 4 presents the experimental results for speaker verification on the handheld-device database. Finally, Section 5 presents a summary of the paper.

II. Universal Compensation (UC) Model

A. Methodology

Denote by Φ_0 the training set containing *clean* training data for a speaker, and denote by $P(X|s, \Phi_0)$ the probability distribution of frame feature vector X associated with speaker s trained on Φ_0. In this study, we assume that each frame vector X consists of N subband feature components: $X = (x_1, x_2, ..., x_N)$, where x_n represents the feature component 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. [26], [27]) for isolating local frequency-band corruption from spreading into the feature components of the other bands.

The first step of the UC method is to multiply the training set Φ_0 by corrupting the clean training data with simulated noise of different characteristics (e.g., white noise at different signal-to-noise ratios (SNRs)). Assume that this leads to augmented training sets Φ_0, Φ_1, ..., Φ_L, where Φ_l denotes the lth training set derived from Φ_0 with the inclusion of a certain noise condition. Then a new probabilistic model for the test frame vector can be formed by combining the probability distributions trained on the individual training sets:

$$P(X|s) = \sum_{l=0}^{L} P(\Phi_l|s) P(X|s, \Phi_l) \qquad (1)$$

where $P(X|s, \Phi_l)$ is the probability distribution of the frame vector trained on set Φ_l and $P(\Phi_l|s)$ is the prior probability for the occurrence of the noise condition represented in Φ_l, for speaker s. Eq. (1) is a multi-condition 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 UC method is to make (1) robust to noise conditions not fully represented in 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 $X = (x_1, x_2, ..., x_N)$ be a test frame vector and $X_{\Phi_l,s} \in X$ be a subset in X containing all the subband components that match the corresponding model components trained in noise condition Φ_l for speaker s. Then, using $X_{\Phi_l,s}$ in place of X as the test vector for each trained noise condition, redefine (1) as

$$P(X|s) = \sum_{l=0}^{L} P(\Phi_l|s) P(X_{\Phi_l,s}|s, \Phi_l) \qquad (2)$$

where $P(X_{\Phi_l,s}|s,\Phi_l)$ is the marginal distribution of the matching subset $X_{\Phi_l,s}$, derived from $P(X|s,\Phi_l)$ with the mismatched subband components ignored to improve mismatch robustness between the test frame X and the trained noise condition Φ_l (i.e. the missing-feature principle). For simplicity, assume independence between the subband components. So the marginal distribution $P(X_{sub}|s,\Phi_l)$ for any subset $X_{sub} \in X$ can be written as

$$P(X_{sub}|s,\Phi_l) = \prod_{x_n \in X_{sub}} P(x_n|s,\Phi_l) \tag{3}$$

where $P(x_n|s,\Phi_l)$ is the probability distribution of the nth subband component for speaker s trained under noise condition l.

Given a test frame X, the matching component subset $X_{\Phi_l,s}$ for each Φ_l and s may be defined as the subset in X that gains maximum probability over the appropriate noise condition and speaker. Such an estimate for $X_{\Phi_l,s}$ is not directly obtainable from (3) by maximizing $P(X_{sub}|s,\Phi_l)$ with respect to X_{sub}. This is because the values of $P(X_{sub}|s,\Phi_l)$ for different sized subsets X_{sub} are of a different order of magnitude and are thus not directly comparable. One way around this is to select the matching subset for noise condition Φ_l and speaker s that produces the highest probability for this noise condition/speaker, *as compared to* the probabilities of the same subset produced for the other noise conditions/speakers $(\Phi_{l'}, s') \neq (\Phi_l, s)$. 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_{sub} as

$$P(s,\Phi_l|X_{sub}) = \frac{P(X_{sub}|s,\Phi_l)P(s,\Phi_l)}{\sum_{s',l'} P(X_{sub}|s',\Phi_{l'})P(s',\Phi_{l'})} \tag{4}$$

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

To incorporate the posterior probability (4) into the model, we first rewrite (1) in terms of $P(s,\Phi_l|X)$,

[1] Dividing the numerator and denominator of (4) by $P(X_{sub}|s,\Phi_l)$ gives

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

Therefore maximizing $P(s,\Phi_l|X_{sub})$ for X_{sub} is equivalent to the maximization of the likelihood ratios $P(X_{sub}|s,\Phi_l)/P(X_{sub}|s',\Phi_{l'})$ for X_{sub}.

i.e., the posterior probabilities of speaker s and noise condition Φ_l given frame vector X:

$$\begin{aligned}
P(X|s) &= \sum_{l=0}^{L} P(\Phi_l|s)P(X|s,\Phi_l) \\
&= \sum_{l=0}^{L} P(\Phi_l|s)\frac{P(X|s,\Phi_l)}{P(X)}P(X) \\
&= \left[\sum_{l=0}^{L} \frac{P(\Phi_l|s)}{P(s,\Phi_l)}P(s,\Phi_l|X)\right]P(X) \\
&= \left[\sum_{l=0}^{L} \frac{1}{P(s)}P(s,\Phi_l|X)\right]P(X) \qquad (5)
\end{aligned}$$

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 subset and assuming an equal prior $P(s)$ for all the speakers, we obtain an operational version of (2) for recognition:

$$P(X|s) \propto \sum_{l=0}^{L} \max_{X_{sub}\in X} P(s,\Phi_l|X_{sub}) \qquad (6)$$

where $P(s,\Phi_l|X_{sub})$ 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 subset can be computationally expensive for large frame vectors X. We simplify the algorithm by approximating each $P(X_{sub}|s,\Phi_l)$ in (4) using the probability for the union of all subsets of the same size as X_{sub}. As such, $P(X_{sub}|s,\Phi_l)$ can be written, with the size of X_{sub} indicated in brackets, as [28]

$$P(X_{sub}(M)|s,\Phi_l) \propto \sum_{\text{all } X'_{sub}(M)\in X} P(X'_{sub}(M)|s,\Phi_l) \qquad (7)$$

where $X_{sub}(M)$ represents a subset with M components ($M \leq N$). Since the sum in (7) includes all subsets, it includes the matching subset that can be assumed to dominate the sum due to the best data-model match. Eq. (7) for $0 < M \leq N$ can be computed efficiently using a recursive algorithm assuming independence between the subband components (i.e. (3)). Note that (7) is not a function of the identity of X_{sub} but only a function of the size of X_{sub} (i.e. M). We therefore effectively turn the maximization in (6) for the identity of the matching subset, of a complexity of $O(2^N)$, to the maximization for the size of the matching subset, $\max_M P(s,\Phi_l|X_{sub}(M))$, of a complexity of $O(N)$, where $P(s,\Phi_l|X_{sub}(M))$ is of a form as (4) with each $P(X_{sub}|s,\Phi_l)$ replaced by the union probability $P(X_{sub}(M)|s,\Phi_l)$. We call $\max_M P(s,\Phi_l|X_{sub}(M))$ the *posterior union model* (PUM), which has been studied previously (e.g. [29]) as a missing-feature method without requiring identity of the noisy data assuming clean data training (i.e. $\Phi_l = \Phi_0$). The UC model (6) is reduced to a PUM with single-condition training (e.g. $L = 0$).

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

$$P(X_1^T|s) = [\prod_{t=1}^{T} P(X_t|s)]^{1/T} \tag{8}$$

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 (8), 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 UC model effectively practices a reconstruction of the test noise condition using a limited number of trained noise conditions. To make the model suitable for a wide range of noises, the multi-condition training sets Φ_1, ..., Φ_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 piece-wise (i.e. band-wise) 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 multi-condition 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 multi-condition 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 unform 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 UC 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 noises at various SNRs are added *acoustically* to clean speech to produce the multi-condition noisy training data. In the showed system, 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 UC 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 by different SNRs - the large number of possibilities makes data collection extremely challenging in conventional multi-condition 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 (which is assumed to be purely additive 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, for example, the loss of high frequency components due to air absorption. 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 [30], may also be captured within the system. The system shown in Fig.1 is used in our experimental studies for speaker identification.

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 distribution 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 UC 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 noise varieties, and on the development of new methods for generating the training data and reducing the model's complexity for the UC model.

The database contains multi-condition training data and test data, both created by using a system illustrated in Fig 1. To create the multi-condition training data for the UC 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 Φ_1, ..., Φ_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 captures 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], [32], [23]); 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 [30] 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 the noise effect and assume the use of a single microphone to record the training and test data (in Section IV we will consider the handset variability for speaker verification on the handheld-device database). The data were recorded in a ordinary office environment, with the use an Electret LEM EMU 4535 microphone, placed about 10 cm from the center of the two loudspeakers 20 cm away from each other. The multi-condition training utterances for the UC 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).

Six different types of real-world noise data were used, respectively, as the test noise source. These were: 1) a jet engine noise, 2) a restaurant noise, 3) a street noise, 4) a polyphonic mobile-phone ring, 5) a pop song with mixed music and voice of a female singer, and 6) a broadcast news segment involving two male speakers with a highway background. 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 test utterances were recorded in the presence of each of the noises at three SNRs: 20, 15 and 10 dB, plus one recoding without noise.

The TIMIT database contains 630 speakers (438 male, 192 female), each speaker contributing 10 utterances and each utterance having an average duration of about 3 seconds. Following the practice in [31], for each speaker, 8 utterances were used for training and the remaining 2 utterances were used for testing. This gives a total of 1260 test utterances across all the 630 speakers. The multi-condition training set for each speaker contained 56 utterances (7 SNRs × 8 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 components derived from the decorrelated log filter-bank amplitudes [33], [34]. Specifically, for each frame a 21-channel mel-scale filter bank was used to obtain 21 log filter-bank amplitudes, denoted by $(a_1, a_2, ..., a_{20}, a_{21})$. 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 $(d_1, d_2, ..., d_{20}) = (a_2 - a_1, a_3 - a_2, ..., a_{21} - a_{20})$. These 20 decorrelated amplitudes were then uniformly grouped into 10 subbands, i.e., $(\{d_1, d_2\}, \{d_3, d_4\}, ..., \{d_{19}, d_{20}\}) \rightarrow (x_1, x_2, ..., x_{10})$, each subband component x_n containing two decorrelated amplitudes corresponding to two consecutive filter-bank channels. These 10 subband components, with the addition of their corresponding first-order delta components, form a 20-component vector $X = (x_1, x_2, ..., x_{10}, \Delta x_1, \Delta x_2, ..., \Delta x_{10})$, of a size of 40 coefficients, for each frame [2].

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

1) BSLN-Cln: a baseline GMM trained on clean data and using all subband components for recognition, with 32 mixtures per speaker;
2) BSLN-Mul: a baseline GMM trained on the simulated multi-condition data and using all subband

[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.

components for recognition, with 128 Gaussian mixtures per speaker;

3) UC: trained on the simulated multi-condition data and focusing recognition on the matching subband components to reduce the training/testing mismatch (i.e. (6), with the maximization implemented by using a PUM as described in that section), with 32, 64 and 128 Gaussian mixtures, respectively, per speaker.

B. Identification Results

Table I presents the identification accuracy obtained by the three 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 UC 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 UC 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 multi-condition model BSLN-Mul, the UC 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 UC and BSLN-Mul. 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, UC 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 UC model offered less significant improvement or no improvement. In our experiments, these noises include the engine noise, restaurant noise and street noise [3]. For these noises, UC and BSLN-Mul 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 UC models, with 32, 64 and 128 mixtures respectively, showed better average performance than the other two systems, indicating the potential of the UC system for dealing with a wider range of test conditions. The relative processing time for the BSLN-Mul with 128 mixtures compared to the UC also with 128 mixtures was about 1:6. This ratio dropped almost linearly to about 1:3 for the UC with 64 mixtures and to about 1:1.5 for the UC with 32 mixtures.

C. Acoustic Noise Addition versus Electronic Noise Addition

In the above experiments the multi-condition training data for the UC 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 multi-condition training data for the UC model. Specifically, in the experiment we assumed that the test data were generated in the same way as above using ANA, but the multi-condition 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

[3] We have conducted an extra experiment that is not included in the paper. In the experiment, we trained a baseline multi-condition 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. This indicates the similarity in characteristics between the three noises and the simulated wide-band noise.

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 effect 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. the Lombard effect) and alter the characteristics of the observed signals (e.g. the handset/channel effect). In Section V we will discuss an possible extension of the UC principle and the training data collection system for modeling new forms of signal distortion.

IV. Speaker Verification Experiments

A. Database and Acoustic Modeling

This section describes further experiments to evaluate the UC model with the use of real-world application data. A handheld-device database [35], designed for speaker verification with limited enrollment data, was used in the experiments (which extend previous results reported in [36]). The database was collected 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 4 times for training and 4 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 4 times in each given background/micophone 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 multi-condition training sets Φ_1, ..., Φ_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 cutoff frequency for the low-pass filter. The latter simulates the weakening high-frequency components for the noise, as may be seen in Fig. 5, due to the loss of the high-frequency components for the relatively distant noise sources by air absorption. In the following, we first present the experimental results for the separate use of the wide-band noise and the narrow-band noise, with a 3dB cutoff frequency of 800 Hz, for training the models. We have tested other cutoff frequencies within the range 700–2000Hz for the narrow-band training noise and found that they offered similar performances. Wide-band training noise is not the best choice for this database with relatively weak high-frequency noise components. However, we have seen in Section III that wide-band training noise is needed for dealing with nearby noise sources with significant high-frequency components. In the final part of this experiment we demonstrate a model built upon the mixed wide-band and narrow-band training noise, 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 8-state HMM, with each state in each condition (i.e. $P(X|s, \Phi_l)$, which now models the observation distribution in state s within a speaker's HMM) being modeled by 2 diagonal-Gaussian mixtures. Additionally, 3 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 feature components (similar to cepstral mean subtraction).

We implemented three systems all based on the same 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 multi-condition data;
3) UC: trained on the simulated multi-condition data.

Two cases were further considered for UC and BSLN-Mul: (a) the use of wide-band noise and (b) the use of narrow-band noise to generate the multi-condition 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 UC and BSLN-Mul trained using narrow-band noise (NB) and wide-band noise (WB) respectively, and for BSLN-Cln. 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 multi-conditionally trained, narrow-band noise based models UC (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. By ignoring some of the mismated data, UC improved the situation, 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), UC obtained lower EERs than the other systems assuming the same information about the test condition.

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). UC 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 UC model by focusing on the matching subbands. As seen, UC (WB) trained on the less matched wide-band noise performed similarly to the BSLN-Mul (NB) trained on the better matched narrow-band noise, in terms of the EER. UC (NB/WB) reduced the EER by 23.4/34.8% as compared to the corresponding BSLN-Mul (NB/WB).

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, UC offered improved performance over both BSLN-Cln and BSLN-Mul. Compared to BSLN-Cln, UC (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, UC (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 UC 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 UC 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 subband (due to the compensation) and from the remaining little corrupted subband (through matched clean subbands between the model and data), and therefore obtain more information, i.e. a larger subset $X_{\Phi_l,s}$ 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 subband of the test signal would have to be ignored to reduce the model-data mismatch, resulting in a loss of information. Without assuming the 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 corruption. In the following we show an example by combining the two UC models described above, one trained on the narrow-band noisy data and the other on the wide-band noisy data, to form a new UC 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 UC (NB). As can be seen, the combined model improved over the wide-band noise based UC (WB), and performed similarly to UC (NB) while retaining the potential of UC (WB) for dealing with wide-band noise corruption. The EERs for the combined model are included in Table II.

Multi-condition 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, multi-condition 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 multi-condition training with a missing-feature model, as evident by the performance differences between

UC and BSLN-Mul.

V. Summary

This paper investigated the problem of speaker recognition in noisy conditions assuming absence of information of the noise. A method, namely universal compensation (UC), was proposed. The UC technique combines multi-condition training and the missing-feature method to model noises with unknown temporal-spectral characteristics. Multi-condition training is conducted using simulated noisy data of simple noise characteristics, providing a coarse compensation for the noise, and the missing-feature method refines the compensation by ignoring noise variations outside the given training conditions, thereby accommodating training and testing mismatch.

We studied the UC model for both speaker identification and speaker verification. The research is focused on new methods for creating multi-condition training data to model realistic 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 UC model as a usual GMM. Two databases were used to evaluate the UC algorithm. The first was a noisy TIMIT database obtained by re-recording the data in various controlled noise conditions, used for an experimental development of the UC model with a focus on the noise varieties. The second was a handheld-device database collected in realistic noisy conditions, used to further validate the UC model by test on the real-world data. Experiments on both databases have shown improved noise robustness for the new UC 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 realistic noisy speech. Acoustic noise addition is made feasible in the UC 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 UC principle to model new forms of signal distortion, e.g. handset variability 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. The missing-feature method will be used to deemphasize the mismatches while exploring the matches arising from the quantized data.

REFERENCES

[1] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Processing, vol. 2, pp. 639-643, Oct. 1994.

[2] R. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition - a feature-based approach," IEEE Signal Processing Magazine, pp. 58-71, Sep. 1996.

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

[4] 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.

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

[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in Proc. A Speaker Odyssey - the Speaker Recognition Workshop, Crete, Greece, 2001.

[7] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in Proc. ICASSP02, Orlando, FL, 2002, pp. 681-684.

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

[9] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data,", in Proc. ICASSP'2003, Hong Kong, China, 2003.

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

[11] H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, "Robust text-independent speaker identification over telephone channels," IEEE Trans. Speech Audio Processing, vol. 7, pp. 554-568, Sep. 1999.

[12] R. Teunen, B. Shahshahani and L. P. Heck, "A model-based transformational approach to robust speaker recognition," in Proc. ICSLP'2000, Beijing, China, 2000.

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

[14] 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.

[15] G. R. Doddington, et al., "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective", Speech Commun., vol. 31, pp. 225-254, 2000.

[16] 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.

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

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

[19] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in Proc. ICASSP'2001, Salt Lake City, UT, 2003.

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

[21] 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.

[22] 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.

[23] 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.

[24] J. Ming, “Universal compensation – an approach to noisy speech recognition assunming no knowledge of noise,” in Proc. ICASSP’2004, Montreal, Canada, 2004, pp. I.961-I.964.

[25] J. Ming, D. Stewart and S. Vaseghi, “Speaker identification in unknown noisy conditions - a universal compensation approach,” in Proc. ICASSP’2005, Philadelphia, PA, 2005.

[26] 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.

[27] H. Hermansky, S. Tibrewala and M. Pavel, “Towards ASR on partially corrupted speech”, in Proc. ICSLP’96, Philadelpia, PA, 1996, pp. 462-465.

[28] J. Ming, P. Jancovic, and F. J. Smith, “Robust speech recognition using probabilistic union models,” IEEE Trans. Speech Audio Processing, vol. 10, pp.403-414, Sep. 2002.

[29] J. Ming and F. J. Smith, “A posterior union model for improved robust speech recognition in nonstationary noise,” in Proc. ICASSP’2003, Hong Kong, China, 2003, pp. 420-423.

[30] D. A. Reynolds, “HTIMIT and LLHDB: speech corpora for the study of handset transducer effects,” in Proc. ICASSP’97, Munich, Germany, 1997.

[31] D. A. Reynolds, “Speaker idenitifcation and verification using Gaussian mixture speaker models,” Speech Commun., vol. 17, pp. 91-108, 1995.

[32] K. P. Markov and S. Nakagawa, “Text-indenpendent speaker recognition using non-linear frame likelihood transformation,” Speech Commun., vol. 24, pp. 193-209, 1998.

[33] 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.

[34] K. K. Paliwal, “Decorrelated and liftered filter-bank energies for robust speech recognition,” in Proc. Eurospeech’99, Budapest, Hungary, 1999, pp. 85-88.

[35] R. Woo, *Exploration of small enrollment speaker verification on handheld devices*, M. Eng. Thesis, MIT Deparment of Electrical Engineering and Computer Science, 2005.

[36] J. Ming, T. J. Hazen, and J. R. Glass, “Speaker verification over handheld devices with realistic noisy speech data,” subbmitted to ICASSP’2006.

[37] L. Deng, A. Acero, M. Plumpe and X.-D. Huang, “Large-vocabulary speech recognition under adverse acoustic environments,” in Proc. ICSLP’2000, Beijing, China, 2000, pp. 806-809.

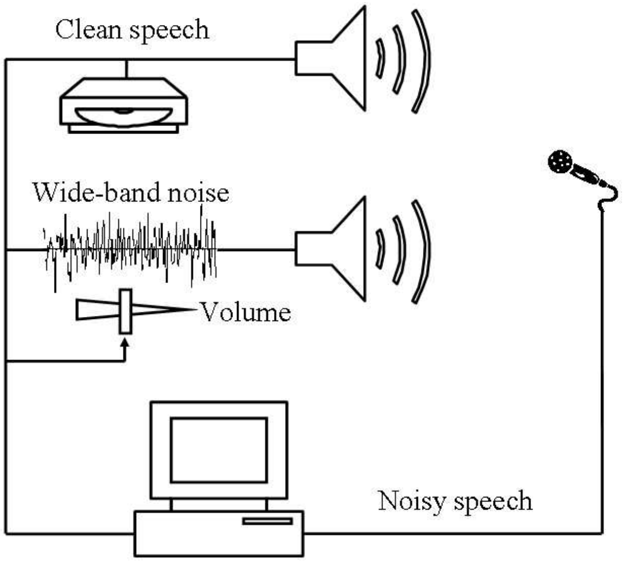

Fig. 1. Illustration of the system used to generate multi-condition training data for the UC model, with wide-band noise of different volumes added acoustically to the clean training data. This system is also used in the experiments to produce noisy test data, by replacing the wide-band noise source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

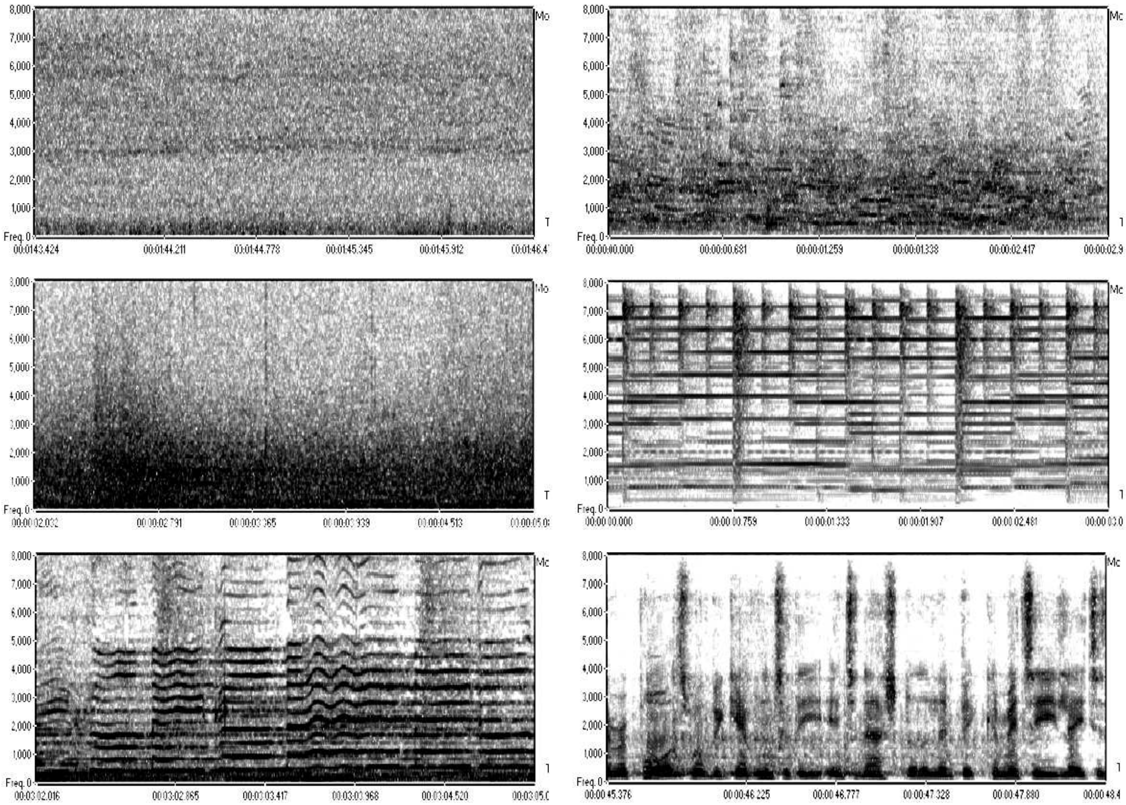

Fig. 2. Noises used in identification experiments, showing the spectra over a period of about three seconds. From left to right, top to bottom: jet engine, restaurant, street, mobile-phone ring, pop song, broadcast news.

TABLE I

Identification accuracy (%) for the universal compensation model (UC) and baseline multi-condition model (BSLN-Mul) trained using simulated, acoustically mixed multi-condition data at seven different SNRs, and for the baseline model trained using clean data (BSLN-Cln). The number associated with each model indicates the number of Gaussian mixtures in the model

Noise	SNR	UC			BSLN-Mul	BSLN-Cln
	(dB)	32	64	128	128	32
Clean		90.64	94.84	96.51	95.79	98.41
Engine	20	83.81	87.06	88.89	86.35	62.46
	15	78.26	81.75	81.59	77.62	29.05
	10	51.27	52.30	51.35	53.57	7.78
Restaurant	20	85.87	91.27	93.89	94.44	93.10
	15	80.56	85.95	88.33	87.46	78.97
	10	67.54	73.25	75.08	67.70	43.57
Street	20	86.75	91.27	92.86	94.29	91.83
	15	79.76	85.08	86.51	86.83	70.32
	10	61.11	63.57	64.05	68.17	34.60
Mobile phone ring	20	73.57	80.64	84.68	68.02	56.90
	15	63.65	72.30	76.35	46.90	34.05
	10	48.10	57.38	62.46	26.43	15.56
Pop song	20	87.54	92.22	93.41	86.19	88.57
	15	78.26	85.71	88.07	64.44	66.98
	10	58.49	64.21	67.70	33.65	30.87
Broadcast news	20	87.22	92.54	93.89	82.78	84.92
	15	79.05	86.03	88.97	59.84	61.75
	10	57.87	66.75	70.00	27.62	26.19

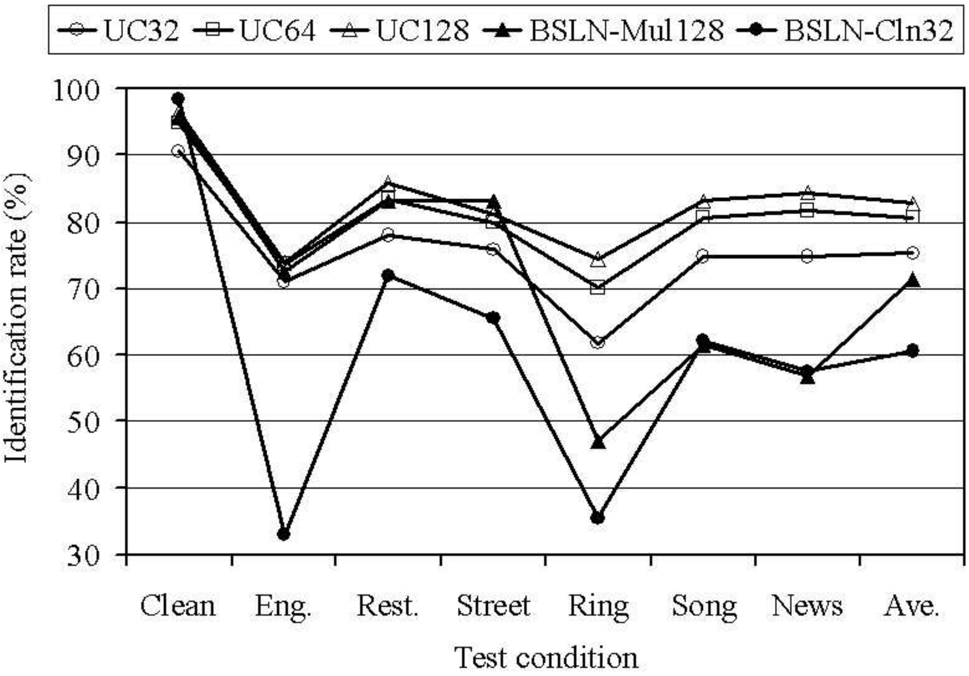

Fig. 3. Identification accuracy in clean and six noisy conditions averaged over SNRs between 10–20 dB, and the overall average accuracy across all the conditions, for UC and BSLN-Mul trained using simulated, acoustically mixed multi-condition data at seven different SNRs, and for BSLN-Cln trained using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

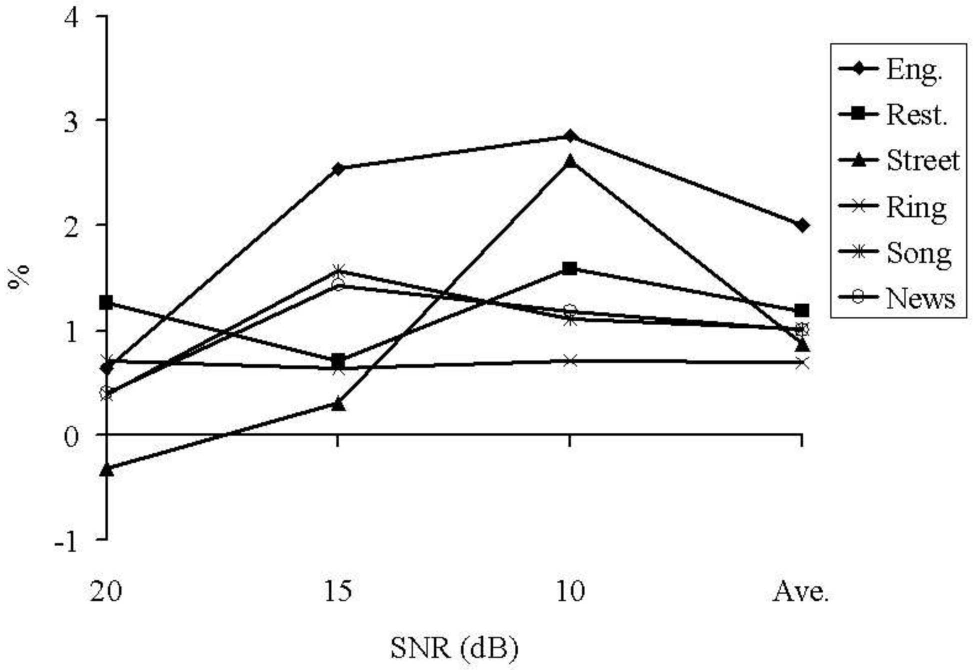

Fig. 4. Absolute improvement in identification accuracy by the UC model trained using multi-condition data with acoustically added noise, compared to a UC model trained using the data with electronically added noise, for test data with acoustically added noise. Both UC models used 128 Gaussian mixtures per speaker.

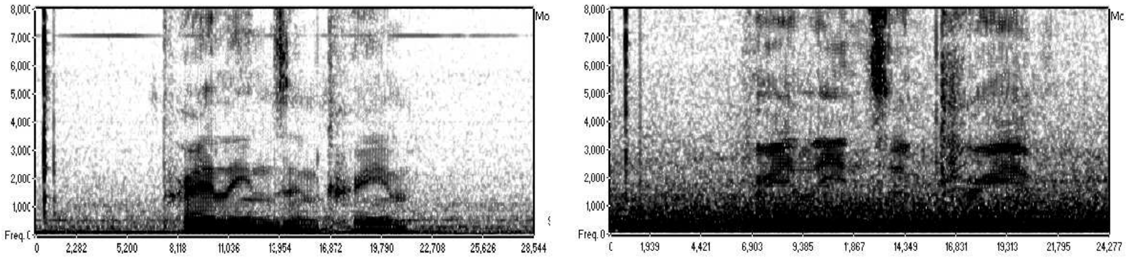

Fig. 5. Spectra of utterances in office (left) and street intersection (right), recorded using the internal microphone.

TABLE II

Equal error rates (%) for UC and BSLN-Mul trained using simulated narrow-band noise (NB), wide-band noise (WB) and combination (NB+WB) at ten different SNRs, and for BSLN-Cln trained using clean data (Index: O–office, S–street intersection, H–headset, I–internal microphone)

Training-Testing	UC			BSLN-Mul		BSLN-Cln
condition	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

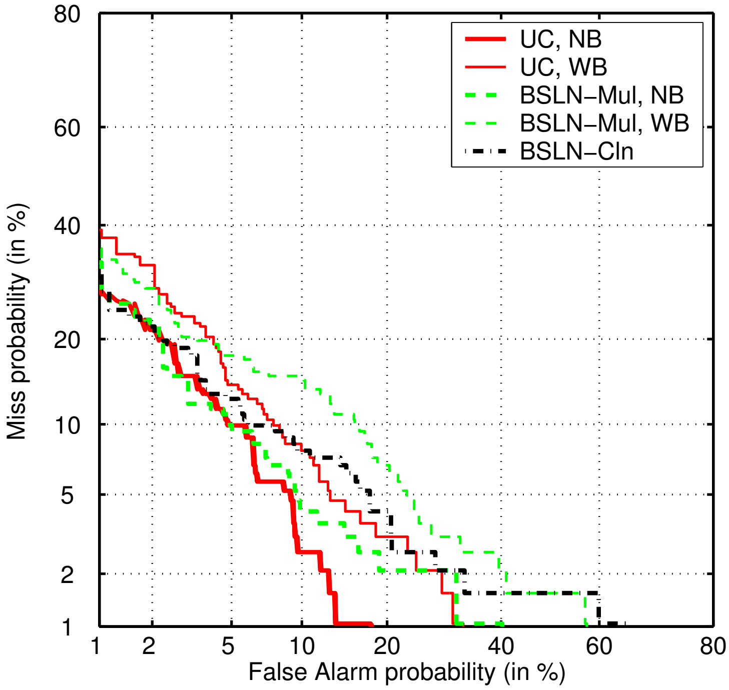

Fig. 6. DET curves in matched training and testing: office/heaset, for UC and BSLN-Mul trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for BSLN-Cln trained using clean data.

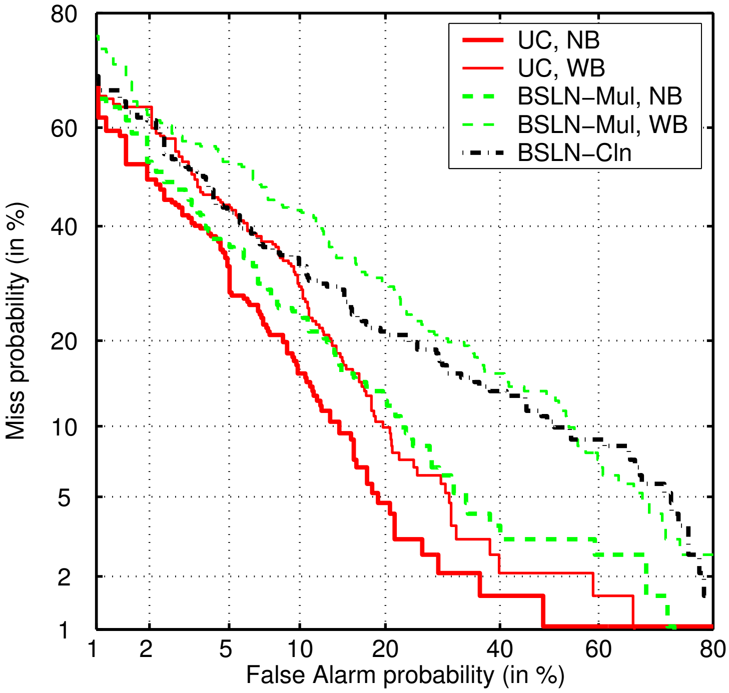

Fig. 7. DET curves with mismatch in environments: training–office, testing–street intersection, both using internal microphone, for UC and BSLN-Mul trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for BSLN-Cln trained using clean data.

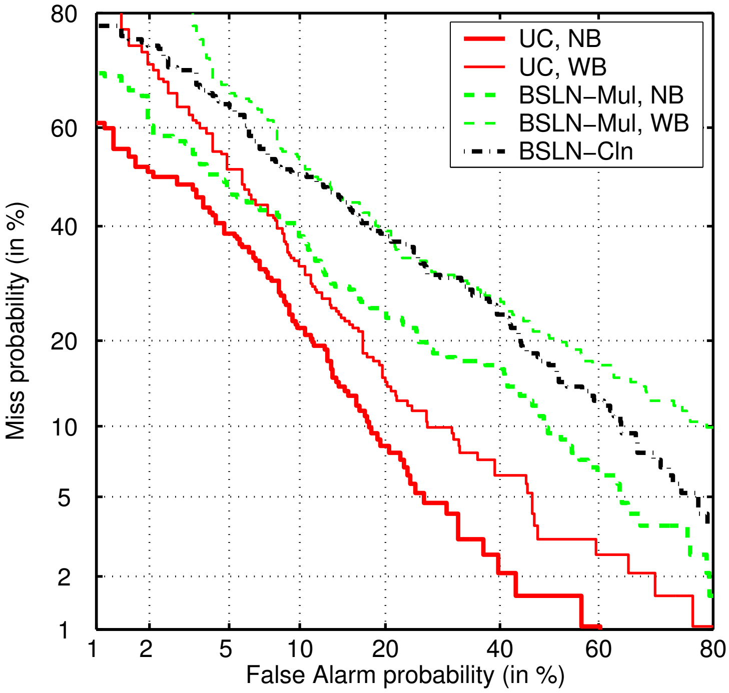

Fig. 8. DET curves with mismatch in both environments and microphones: training–office/internal microphone, testing–street intersection/headset, for UC and BSLN-Mul trained using simulated narrow-band noise (NB) and wide-band noise (WB) at ten different SNRs, and for BSLN-Cln trained using clean data.

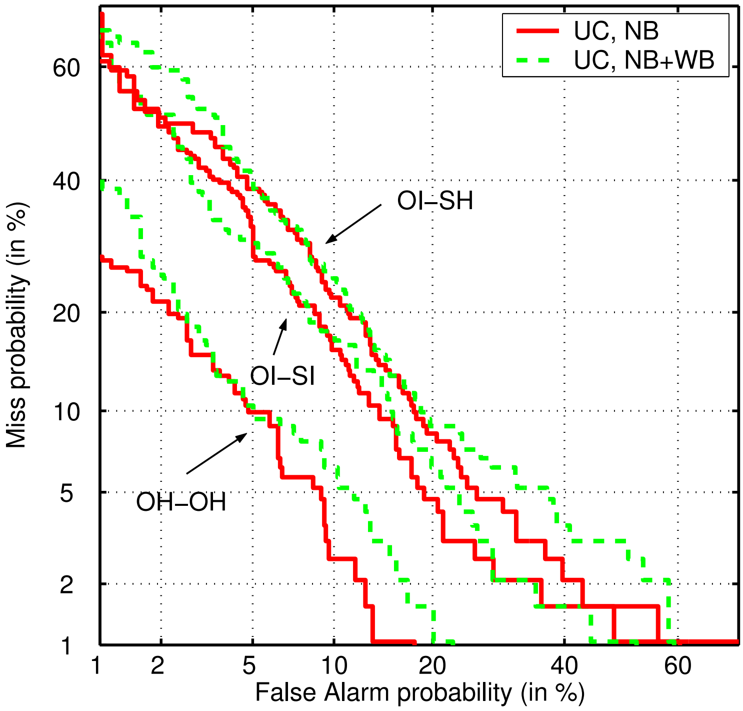

Fig. 9. Comparison between the UC models trained using simulated narrow-band noise (NB) and mixed narrow-band noise and wide-band noise (NB+WB), for different training–testing environment/microphone conditions (Index: O–office, S–street intersection, H–headset, I–internal microphone)