A Comparative Study of Methods for Handheld Speaker Verification in Realistic Noisy Conditions

Ji Ming[†], *Timothy J. Hazen*[‡], *James R. Glass*[‡]

[†]School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[‡]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
j.ming@qub.ac.uk; hazen/jrg@csail.mit.edu

Abstract

A number of methods have been studied and compared for their robustness for speaker verification using noisy speech samples. The methods include Wiener filtering, noise compensation, missing-feature technique, universal compensation, and their combinations. Strategies for combining different techniques are investigated, as a means of further improving noise robustness. A handheld-device database, collected in realistic conditions with noise corruption and transducer mismatch between training and testing, is used in the study. The various techniques and proposed combinations are compared within the same feature and model framework for characterizing the speakers. The experimental results indicate that: 1) usual noise filtering and noise compensation provided very limited robustness to noise corruption, and 2) the proposed technique combinations offered significantly improved noise robustness.

1. Introduction

This paper investigates speaker verification in noisy conditions, assuming that speech signals are corrupted by environmental noise but the characteristics of the noise source are not known *a priori*. This research is motivated in part by the potential application of speaker recognition technologies on handheld devices. While the technologies promise an additional biometric layer of security to protect the user, the practical implementation of such systems faces many challenges, with handset transducer mismatch and environmental noise being two of the most prominent. Recently, much research has been conducted towards reducing the transducer/channel effect (see, for example, [1]–[6]). The present study is focused on the noise issue. Due to the mobile nature of the handheld systems, the acoustic environments and hence 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.

To date, research has targeted the impact of environmental noise through filtering techniques such as spectral subtraction or Kalman filtering [7], [8]. Other techniques rely on a statistical model of the noise, for example, parallel model combination (PMC) [9], [10], or on the use of microphone arrays [11], [12]. Recent studies on the missing-feature method have shown improved robustness for speech data subjected to partial noise corruption (e.g., [13], [14]).

Without assuming a prior knowledge of the noise source, there may be two different approaches to achieving noise robustness: 1) obtaining a noise estimate from the given test signal, and then using the estimate to form a filter for noise removal or to update the acoustic models for noise-effect compensation; 2) building robust acoustic models with inherent robustness to noise corruption. In this paper, we consider examples for both approaches, and furthermore, for their combinations. Specifically, we investigate speaker verification on handheld devices using a database recorded in real-world noisy conditions. We study and compare the robustness of Wiener filtering, noise compensation, missing-feature method, universal compensation, and their combinations. While Wiener filtering and noise compensation are examples of approaches that require an estimate of the noise characteristics, the missing-feature and universal-compensation methods studied in the paper are examples of approaches that do not require information about the noise. Strategies for combining different techniques are investigated, as a focus of the research towards improved noise robustness. The various methods and proposed combinations are compared within the same framework for acoustic modeling. The experimental results show the superiority of the combined techniques to the individual techniques, due to the weakened assumptions and hence enhanced capabilities for modeling real-world noisy speech. This research extends our previous work [15], [16] by focusing on the combination of different modeling techniques for potential performance improvement.

2. Database, Acoustic Modeling and Baseline System

2.1. Database

A handheld-device database [17], designed for speaker verification with limited enrollment data, is used in the study. The database is 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 is used in the experiments. There are 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 are 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 are grouped to form the impostor trials for that enrolled speaker. Then, in each test, there are 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.

Data used in the experiments are recorded in two dif-

1-4244-0472-X/06/$20.00 © 2006 IEEE

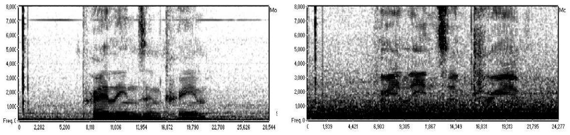

Figure 1: Spectra of utterances in office (left) and street intersection (right), recorded using the internal microphone.

ferent environments: office (with a low level of background noise) and street intersection (with a higher level of background noise). Fig. 1 shows the typical characteristics of the environments. We assume that the speaker models are trained based on the office data and tested in matched and mismatched conditions without assuming prior information about the test environments/microphones.

2.2. Acoustic Modeling

The noise problem may be tackled at different stages of acoustic modeling, including signal preprocessing, feature computation and match-score computation. The paper is focused on the robust techniques applied to the first and third stages. To make these techniques comparable, the same feature and model structures are used for different techniques, such that any observed improvement in recognition performance would be mainly attributable to the improved robustness for the technique being considered.

The speech signal, sampled at 16 kHz, is divided into frames of 20 ms at a frame period of 10 ms. Each frame is modeled by a feature vector characterizing the spectral characteristics of the frame. The frame vector is calculated by first applying a 512-point FFT to the Hamming-windowed frame samples. The resulting 257 FFT magnitudes are then converted to log scale and passed to a 21-channel mel-warped filter bank to obtain 21 log filter-bank energies. The final frame vector is obtained by decorrelating the filter-bank energies using a high-pass filter $H(z) = 1 - z^{-1}$. This results in 20 decorrelated log filter-bank energies (DLFBE), denoted by

$$\begin{aligned} D &= (d_1, d_2, ..., d_{20}) \\ &= (a_2 - a_1, a_3 - a_2, ..., a_{21} - a_{20}) \end{aligned}$$

where a_i stands for the ith log filter-bank energy. DLFBE were studied in [18] and further studied in [19], as an alternative to MFCC (mel-frequency cepstral coefficients) for speech recognition with potentially comparable performance and with less computation than DCT. In this paper, we use D added with its first-order derivative as the frame vector for both the full-feature model (which uses all the feature components for recognition) and the missing-feature model. Denote this 40-component frame vector by

$$X = (d_1, d_2, ..., d_{20}, \Delta d_1, \Delta d_2, ..., \Delta d_{20})$$

where d_n and Δd_n represent the nth static and delta coefficients, respectively. The missing-feature model to be studied is a subband-based model towards exploiting clean frequency-bands while ignoring noisy frequency-bands. The subband features can be formed conveniently from X without extra computation. For example, X can be viewed as a vector consisting of 20 independent subbands, with each subband corresponding to a filter-bank channel. The bandwidth of the subband can be increased conveniently by grouping neighboring subband components together to form a new subband component. In our experiments, we use a 10-subband, 20-stream system obtained by grouping every two consecutive components in X into a new component, i.e.,

$$\begin{gathered} (\{d_1, d_2\}, ..., \{d_{19}, d_{20}\}, \{\Delta d_1, \Delta d_2\}, ..., \{\Delta d_{19}, \Delta d_{20}\}) \\ \rightarrow (x_1, x_2, ..., x_{20}) \end{gathered}$$

where $x_n = \{d_n, d_{n+1}\}$ or $\{\Delta d_n, \Delta d_{n+1}\}$ stands for a new static or delta subband stream containing two static or delta DLFBE modeling two consecutive filter-bank channels. The full-feature model bases recognition on the entire vector X, while the missing-feature model bases recognition on the subband streams x_n assuming least distortion. To reduce the handset transducer effect, the sentence-level mean of X is calculated and removed from X (similar to cepstral-mean subtraction).

In addition to the unified feature structure, we also adopt a unified acoustic model structure for the various techniques to compute match scores for verification. We treat the task as text-dependent speaker verification and model each enrolled speaker using an HMM with eight states for the spoken phrase and three states (tied across all the speakers) for the beginning and ending backgrounds surrounding each utterance. As such, each enrolled speaker is uniquely identified by a particular state subset with the state space consisting of the HMM states of all the enrolled speakers. Denote by $X_1^T = (X_1, X_2, ..., X_T)$ an utterance of T frames, where X_t is the frame vector at time t, and by $S_1^T = (s_1, s_2, ..., s_T)$ the state sequence for X_1^T. The joint probability of X_1^T and S_1^T based on an HMM can be written as

$$\begin{aligned} P(X_1^T, S_1^T) &= P(S_1^T) \prod_{t=1}^{T} P(X_t | s_t) \\ &= P(S_1^T) \prod_{t=1}^{T} \frac{P(X_t | s_t)}{P(X_t)} P(X_t) \\ &= P(S_1^T) P(X_1^T) \prod_{t=1}^{T} \frac{P(s_t | X_t)}{P(s_t)} \qquad (1) \end{aligned}$$

where $P(S_1^T)$ is the Markovian state-sequence probability, $P(X_1^T) = \prod_{t=1}^{T} P(X_t)$, and $P(X|s)$ is the HMM state-emission probability, with $P(s)$ being a prior probability of state s, and $P(s|X)$ being the posterior probability of state s given frame X, defined by

$$P(s|X) = \frac{P(X|s)P(s)}{P(X)} = \frac{P(X|s)P(s)}{\sum_{s'} P(X|s')P(s')} \qquad (2)$$

where the summation in the denominator is over all possible states for frame X. Since $P(X_1^T)$ is not a function of the state index, it can be viewed as a speaker-independent background model. As such, dividing both sides of (1) by $P(X_1^T)$, we obtain a likelihood-ratio function expressed as a function of the posterior probabilities of states, i.e.,

$$\begin{aligned} LR(X_1^T, S_1^T) &= \frac{P(X_1^T, S_1^T)}{P(X_1^T)} \\ &= P(S_1^T) \prod_{t=1}^{T} \frac{P(s_t | X_t)}{P(s_t)} \qquad (3) \end{aligned}$$

Equation (3) may be further simplified by assuming an equal state prior $P(s)$, i.e.,

$$LR(X_1^T, S_1^T) \simeq P(S_1^T) \prod_{t=1}^{T} P(s_t | X_t) \qquad (4)$$

Given an utterance X_1^T and a hypothesized speaker, the verification score VS can be defined as $LR(X_1^T, S_1^T)$ maximized for the state sequence of the HMM of the speaker and normalized for the length of the utterance, i.e.,

$$VS(X_1^T, \hat{S}_1^T) = \max_{S_1^T} \frac{1}{T} \ln LR(X_1^T, S_1^T) \qquad (5)$$

where $\hat{S}_1^T$ denotes the most-likely state sequence for the hypothesized speaker. The maximization in (5) can be computed using the conventional Viterbi algorithm. Equations (4) and (5) are used as the framework for the various techniques to compute speaker scores for verification. The significance of the framework, for unifying the full-feature model and the missing-feature model, will become clear later.

2.3. Test Conditions and Baseline System (BL)

We conduct three tests on the given database. In all the tests, we assume that only the office data are available for training the speaker models. The three tests, indexed by their corresponding enviornment/microphone conditions for training and testing, are described below.

1. OH-OH: both training and testing are conducted in the Office environment with the use of a Headset. This gives matched condition training and testing.
2. OI-SI: training is conducted in the Office environment using the Internal microphone and testing is conducted in the Street-intersection environment also using the Internal microphone. There is a mismatch between the training and testing environments but no mismatch between the microphone types.
3. OI-SH: training is conducted in the Office environment using the Internal microphone and testing is conducted in the Street-intersection environment using a Headset. There are mismatches in both the environments and the microphone types between the training and testing.

We first conducted the tests for a baseline system (BL). The system was a conventional full-feature HMM, using 2 diagonal Gaussian mixtures per state for the eight states modeling the phrase and 16 mixtures per state for the three tied states modeling the backgrounds surrounding the utterance. The system used the feature vector X described in Section 2.2 as the frame vector and used (4) and (5) to compute the speaker score. Fig. 3–5 present the detection-error-tradeoff (DET) curves for the baseline system, for the three test conditions OH-OH, OI-SI, OI-SH described above. Table 1 shows the equal error rates (EERs) produced by the system for the three test conditions. The baseline system accuracy degraded seriously by the noise corruption and microphone mismatch.

3. Wiener Filtering (WF)

A two-stage Wiener filter (WF) [20] is implemented as a preprocessing technique for removing the background noise. The filter is based on an estimate of the noise power spectrum taken at the beginning of each test utterance assuming a period of signal containing only background noise. Twenty frames, or 200 ms, are found suitable for the database for the estimation without requiring an end-point detection. The noise power spectrum is estimated by averaging the FFT power periodograms over the 20 frames. Although a noise tracking algorithm may be further considered for estimating nonstationary noise (e.g., [21]), this is not implemented in our experiments because of the relatively short duration of each test utterance giving few speech-inactive periods within the utterance for noise estimate updating. The WF is used to modify the noisy FFT magnitudes before they are passed to the mel-warped filter bank for calculating the frame vector (Section 2.2). Informal listening tests indicate significantly improved signal-to-noise ratio (SNR) for the filtered noisy signal, along with some mechanical sound effects as usually found with speech enhancement algorithms. Fig. 2 shows an example of the WF effect for increasing the SNR. To reduce the training and testing mismatch, the filter is also applied to the training data. Except for the filtered training/testing data, the HMM structure and the score algorithm remain the same as for the baseline system.

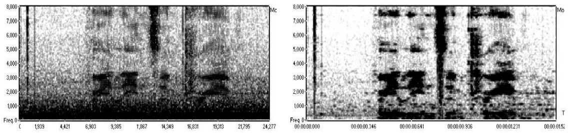

Figure 2: Effect of Wiener filtering, showing the spectrum of a noisy utterance before (left) and after (right) the filtering.

The DET curves for the WF technique for the three test conditions are shown in Fig. 3–5, with the corresponding EERs presented in Table 1. It is seen that the WF improved upon the baseline system for the matched office training and testing (OH-OH), reducing the EER from 8.85% for the baseline system to 5.66%. However, the technique offered very limited performance improvement for the two tests with noise (OI-SI, OI-SH), despite the improved SNR. This may be caused by the inaccuracy of the filter. While removing noise, the filter with inaccurate parameters may also hurt the speech signals. This may not necessarily affect the intelligibility but can change the characteristics of the signal that are critical for speaker recognition. Accurate filtering requires an accurate estimation of the noise characteristics. This can be difficult when there is a lack of data or when the noise characteristics are time varying.

4. Noise Compensation (NC)

Noise compensation (NC) techniques modify the speaker model parameters (e.g., the mean vectors and covariance matrices of the Gaussian mixture model) to match the noise effect on the speech signal. Typical examples of NC include PMC [22], which combines the parameters of a clean speech HMM and a noise HMM to form a new HMM modeling the noisy speech, and multi-condition or multi-style training [23], which builds acoustic models directly on noisy training data matching the test environments. To gain an accurate image of the effectiveness of the method on our database, we consider re-training the speaker models using noise data taken from the test data. As in the implementation of the WF technique, the first 200 ms of each test utterance is taken as the noise data. These are concatenated and added to the waveforms of the office training utterances to form the noisy training data for re-training the speaker models for the appropriate test condition. We believe that this offers a better model than PMC as there are fewer approximations made in forming the noisy model, given the same amount of noise information. The re-trained model takes the same structure as the baseline model and uses the same score algorithm, (4) and (5), as used for the baseline model and the WF model.

Fig. 3–5 show the DET curves for the NC technique for the three test conditions, with the corresponding EERs included in Table 1. Compared to the WF technique, the NC technique performed worse in the OH-OH test, similarly in the OI-SI test, and slightly better in the OI-SH test, with only small improvement over the baseline model for the noisy test conditions (at the price of losing some performance for the office training and testing). A visual examination of the test data indicates sophisticated noise variation in many utterances; the first 200 ms of the signals are not sufficient for capturing these variations.

5. Combining WF, NC with Missing-Feature Technique (WF+MF, NC+MF)

Missing-feature (MF) techniques focus recognition on the least-distorted feature components, thereby reducing the effect of noise on recognition. The techniques are effective given partial noise corruption, a condition that may not be realistically assumed for many real-world applications. This problem may be remedied by combining the MF method with other noise-robust techniques, e.g., WF or NC. WF or NC can be used to deal with the trainable noise component, for example, the slowly-varying or stationary noise component, which is not necessarily partial. This is followed by the MF method that is used to ignore the residual noise leftover by the WF or NC, assuming that this has a partial corruption characteristic. The residual noise may be a combination of the nonstationary noise component difficult to remove by WF or NC, and the distortion caused by inaccurate WF or NC due to inaccurate noise estimation. The combined system thus has the potential of being able to handle full, nonstationary noise corruption.

How to identify the reliable feature parts assuming minimum noise information remains a focus of the research for the MF method. Previous studies have suggested different methods (see, for example, [13], [14], [24]–[26]). In this paper, we study the posterior union model (PUM) [27]. The PUM is applied to frame vector X on a frame-by-frame basis, obtaining an estimate of the reliable feature components within X that maximizes the posterior probability of the associated state $P(s|X)$ as defined in (2). Let $\hat{\chi}$ denote the estimate, then $\hat{\chi} = \arg\max_{\chi \in X} P(s|\chi)$. The maximization can be computed efficiently by approximating the state-emission probability $P(\chi|s)$, for any subset $\chi \in X$, by the probability of the union of all subsets of the same size as χ, i.e. [28],

$$P(\chi|s) \propto \sum_{\text{all } \chi' \in X, \text{size}(\chi')=\text{size}(\chi)} P(\chi'|s) \qquad (6)$$

Since the sum includes all subsets, it includes the least-distorted subset, assuming of the size of χ, that can be assumed to dominate the sum due to the best data-model match. Note that the union probability $P(\chi|s)$ is not a function of the identity of subset χ but only a function of the size of χ. Replacing the state-emission probability in (2) with the union probability (6), we thus turn the maximization for the identity of the reliable subset, $\max_{\chi \in X} P(s|\chi)$, to the maximization for the size of the reliable subset, $\max_{\text{size}(\chi)} P(s|\chi)$, which has a much lower complexity. This is why we call the above model the posterior *union* model.

The PUM can be conveniently incorporated into (4) by replacing $P(s_t|X_t)$ with the state posterior optimized for the feature components, i.e.,

$$LR(X_1^T, S_1^T) \simeq P(S_1^T) \prod_{t=1}^{T} \max_{\chi \in X_t} P(s_t|\chi) \qquad (7)$$

Comparing (4) and (7) indicates a unified score framework for the full-feature model and the missing-feature model. The difference between the two models thus rests only on the utilization of the feature data in deciding the score – an area exploited by the MF method for improving noise robustness.

We repeated the above WF and NC based experiments by using (7) instead of (4) to compute the scores. The results for the combined models, WF+MF, NC+MF, are shown in Fig. 3–5 and Table 1. Both combined models improved significantly upon their previous counterparts, with WF/NC alone, for the noisy test conditions OI-SI, OI-SH. For the office test condition OH-OH, WF+MF performed similarly to WF, and NC+MF performed significantly better than NC. Both combined models improved upon the baseline model with significance. In all the three tests, WF+MF performed better than NC+WF, reducing the average ERR from 19.96% for the baseline model to 11.37%, corresponding to 43% error reduction.

6. Universal Compensation (UC)

Unlike WF and NC which are built upon a noise estimate assuming the availability of training data from the test environments, universal compensation (UC) requires no information about the test noise and hence is suitable for applications without data for noise estimation. UC achieves noise robustness by combining multi-condition training and the missing-feature method. 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. By properly designing the simulated noise data for training, the UC technique has the potential of offering improved robustness for a wide range of noise conditions, e.g., partial-band, full-band, stationary or nonstationary at varying SNRs, without assuming information about the actual noise. Previously we have studied the use of white noise at various SNRs as the training noise, added to the clean training data to form the multi-condition training data for the model [15]. In the present study we consider an alternative, choosing to use the low-pass filtered white noise at various SNRs as the training noise data. The low-pass filtering simulates the high-frequency rolloff characteristics often seen for the realistic noise data, due to the microphone effect, and due to the relatively distant noise sources. The PUM described in Section 5 is used to build the multi-condition model, to exploit the model’s feature-selection ability to focus the recognition on the matching data between the simulated training noise condition and the realistic test noise condition.

Let Φ_0 denote the clean training data set for a speaker (Φ_0 is the office data set in our experiments). The first step of the UC method is to multiply Φ_0 by adding simulated noise to Φ_0 at various SNRs. This leads to multi-condition training sets Φ_1, Φ_2, ..., Φ_L, where Φ_l denotes the lth training set corresponding to a specific SNR. Assume that on each training set Φ_l a speaker model is estimated, which is represented by the HMM state-emission probabilities $P(X|s, \Phi_l)$. The second step of the UC method is to compose $P(X|s, \Phi_l)$ from different sets Φ_l to form a multi-condition model, such that it is capable of accommodating noise varieties, and at the same time capable of ignoring noise variations not matched by the multi-condition training data. The PUM can be extended to implement this. Following (2), define a posterior probability $P(s, \Phi_l|X)$ of state s

and training noise condition Φ_l given frame X:

$$P(s, \Phi_l|X) = \frac{P(X|s, \Phi_l)P(\Phi_l|s)P(s)}{\sum_{s',l'} P(X|s', \Phi_{l'})P(\Phi_{l'}|s')P(s')} \quad (8)$$

where $P(s)$ is a state prior, $P(\Phi_l|s)$ is the the prior probability of the occurrence of the noise condition represented in Φ_l in state s, and the summation in the denominator is over all all possible states and training noise conditions for frame X. A multi-condition model, which produces a state posterior $P(s|X)$ required in (4) for scoring, can be obtained by integrating $P(s, \Phi_l|X)$ over the training noise condition, and by applying the PUM for each training condition to focus on the best-matching test data that maximize $P(s, \Phi_l|X)$, i.e.,

$$P(s|X) = \sum_{l=0}^{L} \max_{\chi \in X} P(s, \Phi_l|\chi) \quad (9)$$

We call (9) the UC model. Comparing (7) and (9) indicates that the PUM is a special case of the UC model with single-condition training (i.e., $L = 0$). As for the PUM, the maximization in (9) for the matching data subset can be turned into a maximization for the size of the matching data subset, and hence with a lower computational complexity, by approximating the state-emission probability $P(\chi|s, \Phi_l)$ in (8), for any subset $\chi \in X$, by the sum $\sum_{\chi'} P(\chi'|s, \Phi_l)$ for all subsets $\chi' \in X$ of the same size as χ, i.e., the probability of the union of all χ'.

In our experiments, we created nine noisy training sets (i.e., $L = 9$) by adding simulated, low-pass filtered white noise to the office training data at nine SNRs from 4 to 20 db (increasing 2 db every step). This gives a total of ten training conditions (including the original office data condition), each condition characterized by a specific SNR. For each speaker, each SNR condition was modeled by an HMM with the same structure as the baseline model as described in Secition 2.3, with eight states with 2 mixtures per state for the spoken phrase and three states with 16 mixtures per state tied across all the speakers for the speech-inactive backgrounds. The state-emission probabilities of these HMMs were combined based on (9) to form the UC model. In computing (8), we assumed a uniform state prior $P(s)$, and a unform noise-condition prior $P(\Phi_l|s)$ assuming no prior knowledge of the structure of the test noise.

The verification results produced by the UC model are presented in Fig. 3–5 and Table 1. Compared to the previous best WF+MF, the UC model offered comparable/better performance for the two noisy test conditions (OI-SI, OI-SH) and a lower average EER over all the three test conditions. Compared to the baseline system, the UC model reduced the average ERR from 19.96% to 10.85%, corresponding to 45.6% error reduction. Note that the UC model achieved these without having assumed knowledge about the test noise.

7. Combining Wiener Filtering and Universal Compensation (WF+UC)

It comes as a natural thought to combine WF and UC as an extension of WF+MF, studied in Section 5, for possible performance improvement. The new WF+UC combination has a potential to improve over WF+MF by removing the assumption that the residual noise leftover by the WF has a partial corruption characteristic, which is required by the WF+MF model for the missing-feature component to function. In the new combination, the residual noise/distortion from the WF can have a full-corruption characteristic, which can be accounted for by the simulated, full-corruption, multi-condition noisy training data, followed by the missing-feature technique (e.g., PUM) to reduce the mismatch of the compensation. It may be assumed that the variance of the residual noise is smaller than that of the original noise. Therefore fewer SNR levels are required in the UC model to account for the residual noise, than required for modeling the original noise. This leads to a smaller UC model.

The idea was examined by experiments on the database. The WF, described in Section 3, was applied to both the office training data and the test data. The office training data, after the filtering, was further corrupted by adding simulated, low-pass filtered white noise to the data at six SNRs from 10 to 20 db (increasing 2 db every step), to form the multi-condition training sets and UC model (9), with $L = 6$. Note that the new UC model had fewer SNR levels than the previous UC model built for the unfiltered noisy test data, described in Section 6. The results for the new combination are shown in Fig. 3–5 and Table 1. The new WF+UC model improved upon the previous WF+MF model in all the three test conditions. The new model also improved over the previous UC model in two of the three test conditions (OH-OH, OI-SI). The new combination produced the lowest average EER among all the seven models studied in the paper, reducing the average EER from 19.96% for the baseline model to 10.19%, corresponding to 48.9% error reduction.

8. Conclusions

This paper investigated different modeling techniques for handheld speaker verification, using a database recorded in realistic noisy conditions. The database provided limited training data for the enrolled speakers, and involved realistic noise and transducer mismatch between training and testing. The modeling techniques being studied include Wiener filtering, noise compensation, missing-feature method, universal compensation, and their combinations. These were studied and compared within the same framework for acoustic featuring, modeling and scoring. Our experimental results on the database indicated that: 1) usual Wiener filtering and noise compensation, based on a noise estimate taken at the beginning of test utterances over a period of signal without speech, showed very limited robustness to noise corruption, and 2) the proposed combined techniques offered significantly improved noise robustness. We studied different combination strategies, including the combination between Wiener filtering/noise compensation and missing-feature method, the combination between multi-condition training and missing-feature method (i.e., universal compensation), and the combination between Wiener filtering and universal compensation (which is effectively the combination of three techniques – Wiener filtering, multi-condition training and missing-feature method). Ideally, the individual component techniques in the combination are complementary to one another. Our further research will be focused on the optimization of the individual component techniques for an optimized combined system.

Acknowledgement

The first author acknowledges the support by the QUB International Exchange Scheme.

9. References

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

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

[3] Barras, C. and Gauvain, J. L., “Feature and score normalization for speaker verification of cellular data,”, in Proc. ICASSP’2003, Hong Kong, China, 2003.

[4] van Vuuren, S., “Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch,” in Proc. ICSLP’96, Philadelpia, PA, 1996, pp. 1788-1791.

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

[6] Xiang, B., Chaudhari, U., Navratil, J., Ramaswamy, G. and Gopinath, R., “Short-time Gaussianization for robust speaker verification,” in Proc. ICASSP’2002, Orlando, FL, 2002, pp. 681-684.

[7] Ortega-Garcia, J. and Gonzalez-Rodriguez, L., “Overview of speaker enhancement techniques for automatic speaker recognition,” in Proc. ICSLP’96, Philadelpia, PA, 1996, pp. 929-932.

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

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

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

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

[12] McCowan, I., Pelecanos, J. and Scridha, S., “Robust speaker recognition using microphone arrays,” in Proc. A Speaker Odyssey - the Speaker Recognition Workshop, Crete, Greece, 2001.

[13] Drygajlo, A. and El-Maliki, M., “Speaker verification in noisy environment with combined spectral subtraction and missing data theory”, in Proc. ICASSP’98, Seattle, WA, 1998, pp. 121-124.

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

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

[16] Ming, J., Hazen, T. J. and Glass, J. R., “Speaker verification over handheld devices with realistic noisy speech data,” to appear in ICASSP’2006, Toulouse, France, 2006.

[17] Woo, R., Park, A. and Hazen, T. J., “The MIT mobile device speaker verification corpus: data collection and preliminary experiments,” IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop, Puerto Rico, 2006.

[18] Nadeu, C., Hernando, J. and Gorricho, M., “On the decorrelation of the filter-bank energies in speech recognition,” in Proc. Eurospeech’95, Madrid, Spain, 1995, pp. 1381-1384.

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

[20] Macho, D., Mauuary, L., Noe, B., Cheng, Y. M., Ealey, D., Jouver, D., Kelleher, H., Pearce, D. and Saadoun, F., “Evaluation of a noise-robust DSR front-end on Aurora databases,” in Proc. ICSLP’2002, Denver, CO, 2002, pp. 17-20.

[21] Martin, R., “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” IEEE Trans. Speech Audio Processing, vol. 9, pp. 504-512, 2001.

[22] Gales, M. J. F. and Young, S. J., “Robust speech recognition in additive and convolutional noise using parallel model combination,” Comput. Speech Lang., vol. 9, pp. 289-307, 1995.

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

[24] Cooke, M., Green, P., Josifovski, L. and Vizinho, A., “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Commun., vol. 34, pp. 267-285, 2001.

[25] Morris, A., Hagen, A., Glotin, H. and Bourlard, H., “Multi-stream adaptive evidence combination for noise robust ASR,” Speech Commun., vol. 34, pp. 25-40, 2001.

[26] Seltzer, M. L., Raj, B. and Stern, R. M., “Classifier-based mask estimate for missing feature method of robust speech recognition,” in Proc. ICSLP’2000, Beijing, China, 2000.

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

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

Table 1: *Equal error rates (%) produced by various modeling techniques including baseline (BL), Wiener filtering (WF), noise compensation (NC), missing-feature (MF), universal compensation (UC) and their combinations, for different environment/microphone conditions: O–office, S–street intersection, H–headset, I–internal microphone.*

Training-Testing condition	Modeling Technique						
	BL	WF	NC	WF+MF	NC+MF	UC	WF+UC
OH - OH	8.85	5.66	10.58	5.89	7.30	6.50	4.58
OI - SI	20.83	19.39	19.28	12.04	16.67	11.98	11.38
OI - SH	30.21	27.62	25.00	16.17	17.72	14.06	14.62
Average	19.96	17.56	18.29	11.37	13.89	10.85	10.19

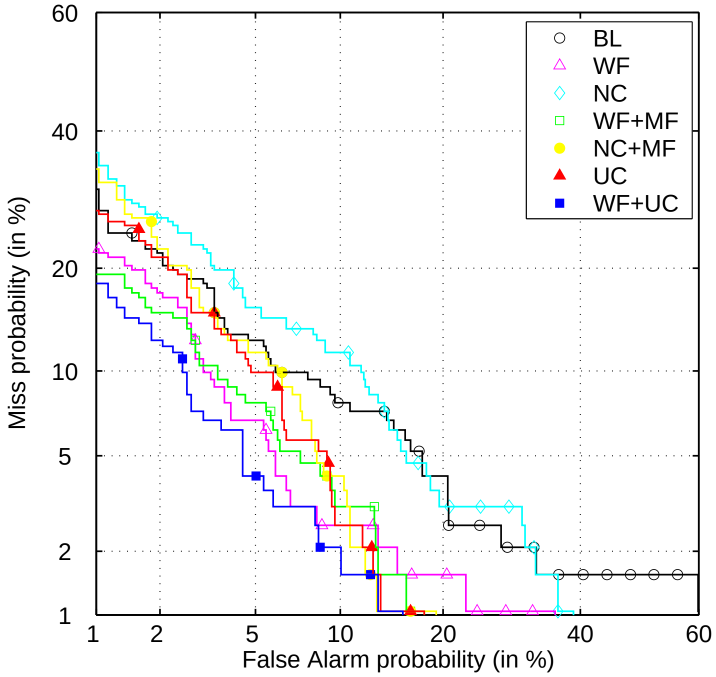

Figure 3: DET curves for matched training and testing OH-OH: training–office/headset, testing–office/headset, for various modeling techniques including baseline (BL), Wiener filtering (WF), noise compensation (NC), missing-feature (MF), universal compensation (UC) and their combinations.

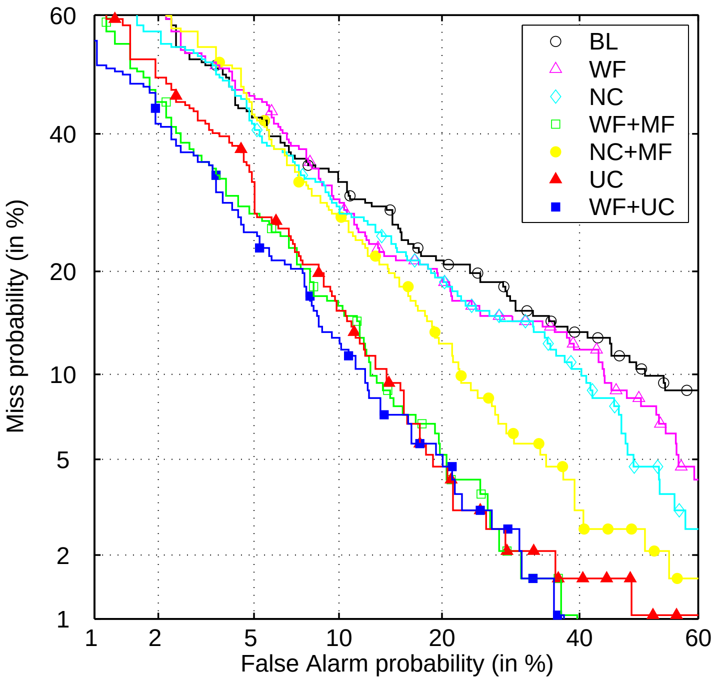

Figure 4: DET curves with mismatched environments OI-SI: training–office/internal microphone, testing–street intersection/internal microphone, for various modeling techniques including baseline (BL), Wiener filtering (WF), noise compensation (NC), missing-feature (MF), universal compensation (UC) and their combinations.

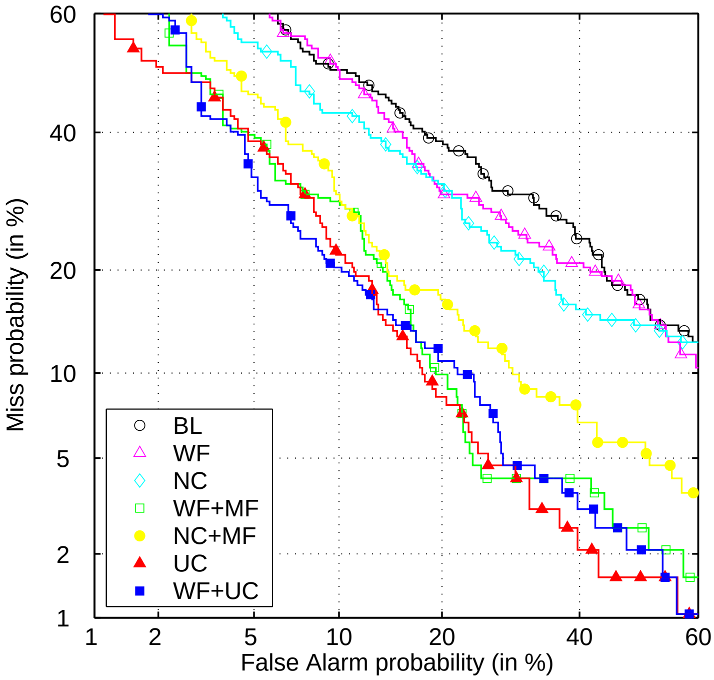

Figure 5: DET curves with mismatched environments and microphone types OI-SH: training–office/internal microphone, testing–street intersection/headset, for various modeling techniques including baseline (BL), Wiener filtering (WF), noise compensation (NC), missing-feature (MF), universal compensation (UC) and their combinations.