SPEAKER VERIFICATION OVER HANDHELD DEVICES WITH REALISTIC NOISY SPEECH DATA

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

[†]School of EECS, Queen's University Belfast, Belfast BT7 1NN, UK
[‡]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

ABSTRACT

We study speaker verification for handheld devices assuming realistic, noisy test conditions and assuming no prior knowledge of the noise characteristics. Data were recorded in office ("quiet") and street intersection ("noisy") environments, with the use of an internal microphone and an external headset. We assume that the speaker models are trained using the office data and tested in matched and mismatched environment/microphone conditions. Two approaches were studied, both built upon a subband feature framework: 1) a posterior union model (PUM) that focuses verification on matching subbands thereby reducing the effect of the training and testing mismatch, and 2) universal compensation (UC) that combines multi-condition training and the PUM to provide robustness to noises of arbitrary temporal-spectral characteristics. Multi-condition training using simulated noise data of different characteristics provides a "coarse" compensation for the noise, and the PUM refines the compensation by ignoring noise variations outside the given training conditions. These two models were compared to baseline systems and have shown improved robustness for realistic noisy speech data.

1. INTRODUCTION

This paper investigates speaker verification in realistic noise conditions, with particular consideration for handheld devices. We tackle a major challenge arising from such devices – mobility, and hence highly time-varying and potentially unknown acoustic environments. Recently, much research has been conducted towards reducing the handset/channel effect. Linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains. Feature compensation includes well-known filtering techniques such as cepstral mean removal or RASTA [1], discriminative feature design [2] and various feature transformations (e.g., [3]). Score-domain compensation includes H-Norm [4], Z-norm [5] and T-Norm [6]. Model-domain compensation includes the speaker-independent variance transformation, and the transformation for synthesizing supplementary speaker models for other channel types from multi-channel training data [7]. Additionally, channel mismatch has been tackled using model adaptation methods, effectively using new data to learn channel characteristics (e.g., [8]).

To date, research has targeted the impact of environmental noise through filtering techniques such as spectral subtraction or Kalman filtering [9], [10], assuming *a priori* knowledge of the noise spectrum. Other techniques rely on a statistical model of the noise, for example, PMC [11], or on the use of microphone arrays [12]. 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., [13], [14]). Missing-feature techniques are effective given partial noise corruption, a condition that may not be realistically assumed for many real-world problems.

This study aims to develop a method that enables the modeling of unknown, time-varying noise corruption without assuming prior knowledge of the noise statistics. A new 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 method involves a combination of the multi-condition training method and the missing-feature method. Multi-condition training, with simulated noisy data of different noise characteristics, 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 the given training conditions, thereby accommodating mismatches between the simulated training noise condition and the realistic test noise condition. In our implementation, the posterior union model (PUM) [15] – a missing-feature model without assuming identity of the corrupted data, was used to estimate the matching data between the model and the test signals. A preliminary study on the UC model for speaker *identification* with synthetic noisy data was reported in [16].

2. METHODOLOGY

2.1. Universal Compensation (UC) Model

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. Assume that each frame vector $X = (x_1, x_2, ..., x_N)$ consisting of N subband feature components, with x_n representing the nth subband component. 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 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:

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

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

This work was supported in part by Intel Corporation. The first author also acknowledges the support by the QUB International Exchange Scheme.

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 noises 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 and $X_{l,s} \in X$ be a subset in X containing all the subband components that match the corresponding model components trained in noise condition l for speaker s. Then, using $X_{l,s}$ in place of X as the test vector for each trained noise condition, redefine (1) as

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

where $P(X_{l,s}|s,\Phi_l)$ is the marginal distribution of the matching subset $X_{l,s}$, 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 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 $X_{l,s}$ for each 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 $X_{l,s}$ 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 normalize $P(X_{sub}|s,\Phi_l)$ using the probabilities for the same subset from all the speakers and noise conditions, and then select the matching subset by maximizing the normalized probability. 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 the subset calculated over all speakers and trained noise conditions, with $P(s,\Phi_l) = P(\Phi_l|s)P(s)$ being a speaker/noise condition prior. Maximizing posterior probability $P(s,\Phi_l|X_{sub})$ for X_{sub} leads to an $X_{l,s}$ estimate that effectively maximizes the likelihood ratios $P(X_{l,s}|s,\Phi_l)/P(X_{l,s}|s',\Phi_{l'})$ for (s,Φ_l) compared to all $(s',\Phi_{l'}) \neq (s,\Phi_l)$.

Rewrite (1) in terms of the posterior probabilities $P(s,\Phi_l|X)$:

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

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

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

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 frames 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 [15]

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

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 [15] 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), can be used directly for speaker verification as well as for identification.

2.2. Generation of Multi-Condition Training Data

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, white noise at consecutive SNRs can be used to corrupt the training data. This spans the full frequency range and a wide amplitude range and therefore allows the expression of sophisticated noise spectral structures. Alternatively, low-pass filtered white noise of different SNRs may be considered for the 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 train the model. Without prior knowledge of the structure of the test noise, a unform prior $P(\Phi_l|s)$ can be used to combine different noise conditions. To limit the size of the model,

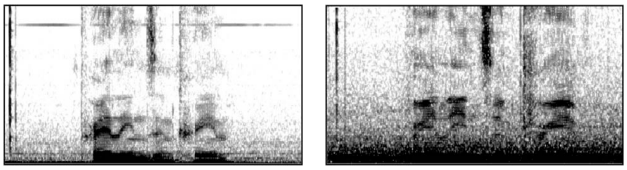

Fig. 1. Spectra of utterances in office (left) and intersection (right)

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.

3. EXPERIMENTS

Experiments were conducted on a handheld-device database, collected at MIT for studying speaker verification in realistic noisy conditions with limited enrollment data [17]. 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.

Data were collected in two different environments: office (with a low level of background noise) and street intersection (with a higher level of background noise), using two different types of microphone: internal (built in the device) and external (a headset). Fig. 1 shows the typical characteristics of the environments. The speaker models were trained based on the office data and tested in matched and mismatched conditions. 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 added low-pass filtered white noise to each training utterance at nine SNRs from 4 to 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, each state each condition (i.e., $P(X|s,\Phi_l)$) modeled by 2 diagonal-Gaussian mixtures. Additionally, 3 states with 16 mixtures per state were used to account for the beginning and ending backgrounds; they were tied across all the speakers.

The speech was divided into frames of 20 ms at a frame rate of 10 ms. Each frame was modeled by a feature vector consisting of subband components derived from the decorrelated log filter-bank amplitudes [18]. Specifically, for each frame a 21-channel mel-scale filter bank was used to obtain 21 log filter-bank amplitudes. These were decorrelated by using a high-pass filter $H(z) = 1 - z^{-1}$ into 20 decorrelated log filter-bank amplitudes. These 20 decorrelated amplitudes were then uniformly grouped into 10 subbands, each subband containing two decorrelated amplitudes corresponding to two consecutive filter-bank channels. These 10 subband components, with the subtraction of the sentence-level mean (similar to cepstral mean removal) and with the addition of their corresponding first-order delta components, form a 20-component vector $X = (x_1, x_2, ..., x_{20})$, of a size of 40 coefficients, for each frame. We implemented four systems all based on the same feature format, and all having the same state-mixture topology as described above:

1. PUM: trained on "clean" (office) data and optimally selecting subband components for recognition
2. UC: trained on the simulated multi-condition data and using PUM to reduce training/testing mismatch
3. BSLN-Cln: a baseline system trained on office data as for PUM and using all subband components for recognition
4. BSLN-Mul: a baseline system trained on multi-condition data as for UC and using all subband components for recognition

Table 1. Closed-set identification rates for enrolled speakers (Index: o–office; s–street intersection; h–headset; i–internal microphone)

Training-Testing	PUM	UC	BSLN-Cln	BSLN-Mul
oh-oh	97.40	97.92	91.15	97.40
oi-si	82.29	89.06	53.12	71.35
oi-sh	67.19	80.21	29.69	56.77

We first compared the four systems assuming matched condition training and testing, both in the office environments with the use of a headset. Fig. 2 presents the DET curves, along with the equal error rate (EER) for each system. 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. By ignoring some of the mismatched data, both PUM and UC performed better than their counterparts, BSLN-Cln and BSLN-Mul. Also, training the models using the simulated multi-condition data showed usefulness for reducing the mismatch, as seen for the better performances obtained by the two multi-conditionally trained models: UC, BSLN-Mul.

Next, we tested the four systems assuming there is training/testing mismatch in environments but not 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. 3 shows the results. Both PUM and UC offered significantly improved performance, reducing the EER by 44.9/42.5% (PUM/UC) as compared to BSLN-Cln, and by 26.6/23.4% as compared to BSLN-Mul. BSLN-Mul improved over BSLN-Cln with the inclusion of the simulated noisy training data. In this case, UC performed similarly to PUM for verification.

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. 4 presents the results. Again, both PUM and UC offered improved performance, reducing the EER by 43.0/53.4% (PUM/UC) as compared to BSLN-Cln, and by 23.2/37.2% as compared to BSLN-Mul. UC outperformed PUM in this case, showing better robustness to the combined mismatch.

Finally, the closed-set identification results for the 48 enrolled speakers are provided in Table 1. Two observations may be drawn: 1) multi-condition training with the simulated noisy data improved the performance, as seen by the performance differences between PUM/UC, and between BSLN-Cln/BSLN-Mul; 2) further improved robustness may be obtainable by combining multi-condition training with a missing-feature model, as evident by the performance difference between UC and BSLN-Mul.

4. CONCLUSIONS

This paper showed that a combination of multi-condition training and the missing-feature theory has the potential to offer improved

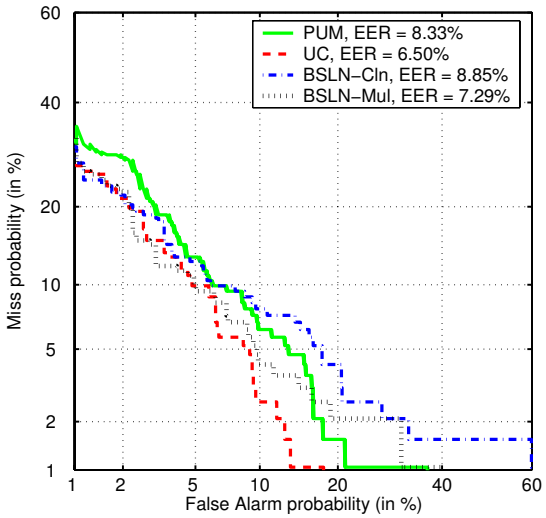

Fig. 2. Performance in matched training and testing: office/headset

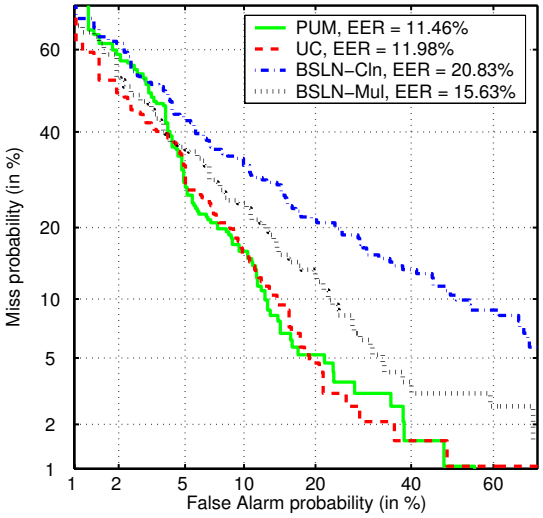

Fig. 3. Performance with mismatch in environments: training – office, testing – street intersection, both using internal microphone.

noise robustness in the absence of information of the noise. The method, namely universal compensation (UC), was tested on a handheld device database collected in realistic noisy conditions for speaker verification. Trained on clean data and simulated, simple noisy data, the UC model showed encouraging robustness to sophisticated realistic noise. Further research will be focused on the design of improved multi-condition training set for the UC model to better model real-world noisy speech data.

5. REFERENCES

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

[2] L. P. Heck, et al., "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," Speech Communication, vol. 31, pp. 181-192, 2000.

[3] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," Speaker Odyssey'01.

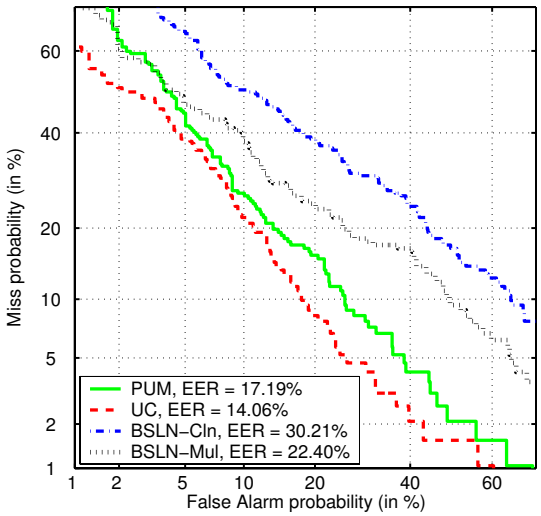

Fig. 4. Performance with mismatch in both environments and microphones: training – office/internal microphone, testing – street-intersection/headset.

[4] 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.

[5] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data,", ICASSP'2003.

[6] 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.

[7] R. Teunen, et al., "A model-based transformational approach to robust speaker recognition," ICSLP'2000.

[8] K. K. Yiu, M. W. Mak and S. Y. Kung, "Environment adaptation for robust speaker verification," Eurospeech'03.

[9] J. Ortega-Garcia, et al., "Overview of speaker enhancement techniques for automatic speaker recognition," ICSLP'96.

[10] Suhadi, et al., " An evaluation of VTS and IMM for speaker verification in noise," Eurospeech'2003, pp. 1669-1672.

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

[12] I. McCowan, J. Pelecanos and S. Scridha, "Robust speaker recognition using microphone arrays," Speaker Odyssey'01.

[13] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory", ICASSP'98, pp. 121-124.

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

[15] Ji Ming and F. J. Smith, "A posterior union model for improved robust speech recognition in nonstationary noise," ICASSP'2003, pp. 420-423.

[16] Ji Ming, D. Stewart and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," ICASSP'2005, pp. 617-620.

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

[18] C. Nadeu, et al., "On the decorrelation of the filter-bank energies in speech recognition," Eurospeech'95, pp. 1381-1384.