EXPLOITING LONG-RANGE TEMPORAL DYNAMICS OF SPEECH FOR NOISE-ROBUST SPEAKER RECOGNITION

Ayeh Jafari, Ramji Srinivasan, Danny Crookes, Ji Ming

Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Belfast BT3 9DT, UK
phone: + (44) 28 90971705, fax: + (44) 28 90974879, email: ajafari01, r.srinivasan, d.crookes, j.ming@qub.ac.uk

ABSTRACT

Temporal dynamics is an important feature of speech that distinguishes speech from noise, as well as distinguishing between different speakers. In this paper, we present an approach to *maximally* extract this feature of speech to improve the robustness against background noise, for text-independent speaker recognition. The new approach identifies and compares the *longest* matching speech segments between the training and test speech to increase noise immunity. Experiments have been conducted on the NIST 2002 SRE database in the presence of various types of noise including fast-varying song and music. The new approach has shown significantly improved performance over conventional noise-robust techniques.

1. INTRODUCTION

In current speaker recognition systems, methods to reduce the influence of background noise include one or combinations of the following: 1) speech enhancement [1, 2, 3], 2) robust acoustic features [4, 5, 6], and 3) noise compensation (e.g., parallel model combination, multicondition model training, and missing-feature decoding) [7, 8, 9]. Most methods are focused on the modeling of *noise*, and are applied within a Gaussian mixture model (GMM) framework in which a GMM is used to model a speaker. A speaker's GMM describes the probability distribution of the speaker's short-time sounds (i.e., frames), but assumes statistical independence between consecutive frames. Therefore, the GMM fails to capture the temporal dynamics of speech, which describes how short-time sounds can be concatenated one to another to form a realistic utterance. Long-range temporal dynamics is one of the most important features of speech which distinguishes speech from non-speech noise, and one speaker's voice from other speakers' voices.

In this paper, we study the problem of improving noise robustness by focusing on the modeling of *speech*, particularly, its long-range temporal dynamics. For text-independent speaker recognition, how to effectively capture long-range temporal dynamics of speech remains a focus of research. Researchers have studied text-constrained speaker recognition, based on common subword or word units between the training and test data [10, 11]. Alternatively, recognition has been based on acoustic segments identified either by phonetic similarity or by minimum distance [12]. Other methods include the use of prosodic features [13] and phonetic refraction, expressed as phone n-gram counts [14].

In this paper, we propose a method to *maximally* extract the temporal dynamics of speech, with the aim of maximizing the noise robustness arising from this distinct feature of speech. We achieve this by identifying and comparing the *longest* matching segments between the test data and training data. Longer speech segments as whole units contain more distinct temporal dynamics, and can be identified more accurately from noise than shorter speech segments. Therefore speaker recognition based on the longest matching segments effectively maximizes noise immunity, and hence reduces the requirement for information about the noise. In the paper, we provide examples to demonstrate that the new approach offers improved robustness over GMM-based recognizers, and robustness against nonstationary or unpredictable noise which is difficult to model with conventional noise modeling approaches. This work is an extension of our previous work [15] from modeling clean speech to noisy speech. In the following, we first introduce the new longest matching segment approach for speaker recognition using clean speech, and then extend the approach to speaker recognition using noisy speech assuming minimal information about the noise.

This work was supported by the UK EPSRC grant EP/G001960.

2. THE LONGEST MATCHING SEGMENT FRAMEWORK FOR SPEAKER RECOGNITION

The longest matching segment (LMS) framework is a new approach for segment-based speaker recognition. It improves speaker discrimination and noise robustness by maximizing the size of speech segments to be compared between the test and training sentences. The speech segments obtained are of arbitrary-length of consecutive frames from speech sentences, which may be of any sound made by a speaker, not limited to a subword unit, and not necessarily phonetically transcribable.

As in the GMM framework, in the LMS framework a speaker is represented through a GMM. However, the LMS framework moves further, by modeling the *full* temporal dynamics in each training speech sentence, and by performing recognition based on the *longest* common speech segments between the training and test data. Let G_λ represent a GMM for speaker λ modeling the probability distribution of the speaker's short-time speech frames x

$$G_\lambda = \{g(x|k,\lambda), w(k|\lambda) : k = 1,2,...,K\} \quad (1)$$

where $g(x|k,\lambda)$ is the k'th Gaussian component and $w(k|\lambda)$ is the corresponding weight. Based on the GMM, the LMS approach further builds a model for each training sentence from the speaker, to capture the full temporal dynamics in the sentence. Let $\mathbf{x} = \{x_i : i = 1,2,...,I_\mathbf{x}\}$ be a training sentence from speaker λ with $I_\mathbf{x}$ frames. This sentence can be represented by a time sequence of the Gaussian indexes, which

address the Gaussian components in G_λ that produce maximum likelihoods for the corresponding frames. This *sentence model* can be expressed as

$$(\mathbf{k}_\mathbf{x}, \lambda) = \{(k_{\mathbf{x},i}, \lambda) : i = 1, 2, ..., I_\mathbf{x}\} \quad (2)$$

where $(k_{\mathbf{x},i}, \lambda)$ indexes a Gaussian $g(x|k_{\mathbf{x},i}, \lambda)$ in G_λ that produces maximum likelihood for frame x_i in the training sentence $\mathbf{x}$. As can be noticed, the above representation shares characteristics with a template. However, it provides a smoother, and hence more robust, representation than templates by modeling each frame using a Gaussian component. In the training stage, we create a model $(\mathbf{k}_\mathbf{x}, \lambda)$ for each training sentence $\mathbf{x}$ for each speaker λ. All these training sentence models for a speaker together form a model for the speaker, used in the recognition.

In recognition, instead of comparing individual frames as in the normal GMM framework, we identify and compare matching segments of consecutive frames between the training and test data, as a means of increasing the speakers' discrimination and noise robustness arising from the temporal dynamics of speech. More specifically, we perform recognition based on the *longest* matching segments between the training and test sentences, aiming to maximize the discrimination/robustness. Let $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$ be a test sentence with T frames, and $\mathbf{y}_{t:\tau} = \{y_\varepsilon : \varepsilon = t, t+1, ..., \tau\}$ be a test segment in $\mathbf{y}$ from frame t to τ. Based on (2), a training segment can be expressed as $(\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \{(k_{\mathbf{x},i}, \lambda) : i = u, u+1, ..., v\}$, which corresponds to the segment from frame u to v in training sentence $\mathbf{x}$ for speaker λ. We compare the two segments, $\mathbf{y}_{t:\tau}$ and $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$, by using the posterior probability $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau})$. Assuming an equal prior P for all the training segments, $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau})$ can be expressed as:

$$\begin{aligned} P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau}) &= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)P}{p(\mathbf{y}_{t:\tau})} \\ &= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)}{\sum_{\lambda'}\sum_{\mathbf{x}'}\sum_{u',v'} p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x}',u':v'}, \lambda') + p(\mathbf{y}_{t:\tau}|\phi)} \end{aligned} \quad (3)$$

In the denominator, the first term corresponds to the likelihood that $\mathbf{y}_{t:\tau}$ matches a training segment, averaged over all the training segments from all the training sentences/speakers; the second term is the likelihood that $\mathbf{y}_{t:\tau}$, *as a whole unit*, is not seen in the training data. This likelihood of unseen test segments can be calculated using a GMM trained with training data from all the speakers [15]. Based on (2), the segment likelihood function $p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ can be written as

$$p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \prod_{\varepsilon=t}^{\tau} g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda) \quad (4)$$

where i_ε is the most-likely time map between the two segments, which maps test frames y_ε to training frames $(k_{\mathbf{x},i_\varepsilon}, \lambda)$, assuming $i_t = u$ and $i_\tau = v$. It can be shown that, when longer $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ and $\mathbf{y}_{t:\tau}$ are matched, larger posterior probabilities $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau})$ are obtained [15]. Therefore, the recognition problem can be expressed as to find the speaker λ which maximizes the sentence score $\Gamma(\lambda; \mathbf{y})$:

$$\Gamma(\lambda; \mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \max_{\tau} \max_{\mathbf{k}_{\mathbf{x},u:v} \in \lambda} \log P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau}) \quad (5)$$

At each test frame time t, the expression seeks to find the *longest* matching training and test segments by jointly maximizing the posterior probability over all training segments and all possible test segment lengths (i.e., τ).

3. EXTENSION TO NOISY SPEECH

Many current studies for noise-robust speaker recognition focus on the modeling of *noise*, within the GMM-based speaker models. In this paper, we focus on the modeling of long-range temporal dynamics of *speech*, and the combination with models of noise, for improved noise robustness arising from the new speech/speaker model. Specifically, we extend the above LMS approach and consider the identification and comparison of the longest matching speech segments between the training data and *noisy* test data. Since longer speech segments, when treated as whole units, can be identified more accurately from noise than individual frames, recognition based on the longest matching segments increases the noise immunity.

In the above LMS framework, a noisy test segment $\mathbf{y}_{t:\tau}$ is compared directly to a clean training segment $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ using the posterior probability $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau})$. We can make this comparison more robust to the noise in $\mathbf{y}_{t:\tau}$ by combining noise compensation. In this paper, we use a missing-feature based approach which assumes minimal information about the noise. In the training stage, we simulate the test noise by adding variable forms of noise to the clean training sentences. As such, we compare the noisy test sentence with the *noisy* training sentences to reduce the noise-caused mismatch. Let ω_n, $n = 1, 2, ..., N$, represent N training noise conditions. Thus, each Gaussian component $g(x|k, \lambda)$ in the speaker's GMM G_λ, (1), can be expanded to a set of Gaussian components $g(x|k, \lambda, \omega_n)$, $n = 0, 1, ..., N$, where $g(x|k, \lambda, \omega_n)$ is estimated using the frames corresponding to $g(x|k, \lambda)$ but corrupted at noise condition ω_n, with ω_0 denoting the noise-free condition. Thus, the time sequence (2) can be extended to model a training sentence corrupted at N different noise conditions. This model addresses a sequence of Gaussian sets with each set, $(k_{\mathbf{x},i}, \lambda, \omega_n) : n = 0, 1, ..., N$, modeling a frame in the training sentence $\mathbf{x}$ from speaker λ corrupted at variable noise conditions ω_0 through ω_N. This can be expressed as

$$(\mathbf{k}_\mathbf{x}, \lambda) = \{(k_{\mathbf{x},i}, \lambda, \omega_n) : n = 0, 1, ..., N; i = 1, 2, ..., I_\mathbf{x}\} \quad (6)$$

Based on (6), the segment likelihood function (4) can be rewritten as

$$p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \prod_{\varepsilon=t}^{\tau} p(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda) \quad (7)$$

where

$$p(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda) = \sum_{n=0}^{N} g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n)P(\omega_n) \quad (8)$$

is a multicondition model of the likelihood of test frame y_ε with a prior probability $P(\omega_n)$ for condition ω_n (assumed to be a uniform distribution in the paper). This new model should improve upon the clean-condition model (4) by offering robustness to the variable noise conditions seen in the training.

In the recognition stage, we can further extend the noise robustness beyond the training conditions by deemphasizing the local frequency-band mismatches between the training and testing noise conditions. For this, we represent each speech frame y_ε using an F-subband vector $y_\varepsilon = (y_{\varepsilon,1}, y_{\varepsilon,2}, ..., y_{\varepsilon,F})$, where $y_{\varepsilon,f}$ is the feature for the fth subband. At each training noise condition ω_n, we assume that y_ε can be divided into two subsets. One subset, $y_\varepsilon(\omega_n)$, includes the subband features that are matched by the training noise condition ω_n; the other subset, the complement $\tilde{y}_\varepsilon(\omega_n)$, includes the rest of the subband features that are mismatched by the training noise condition. Improved robustness can be obtained by computing the frame likelihood (8) by replacing $g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n)$ based on the full feature set with $g(y_\varepsilon(\omega_n)|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n)$ based on the matched feature set for each condition (i.e., the missing-feature theory):

$$g(y_\varepsilon(\omega_n)|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n) = \prod_{y_{\varepsilon,f} \in y_\varepsilon(\omega_n)} g(y_{\varepsilon,f}|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n) \quad (9)$$

where $g(y_{\varepsilon,f}|k_{\mathbf{x},i_\varepsilon}, \lambda, \omega_n)$ is the likelihood of the f'th subband in frame y_ε, and we assume independence between the subbands.

However, (8) is not in a form suitable for estimating the optimal feature sets $y_\varepsilon(\omega_n)$, as values of $g(y_\varepsilon(\omega_n)|k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)$ for different sized $y_\varepsilon(\omega_n)$ are in different order of magnitude and are thus incomparable. We can make the frame likelihood $p(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda)$ effectively comparable for different feature subsets by expressing each $g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)P(\omega_n)$ in (8) through a posterior probability. We use the expression

$$\begin{aligned} g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)P(\omega_n) &= \frac{g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)P(\omega_n)}{p(y_\varepsilon)} p(y_\varepsilon) \\ &= P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y_\varepsilon)p(y_\varepsilon) \end{aligned} \quad (10)$$

In (10), the last term $p(y_\varepsilon)$ is not a function of the matching training frame and hence can be ignored from computation; $P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y_\varepsilon)$ is the posterior probability of the training frame $(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)$ given test frame y_ε, which can be expressed as

$$P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y_\varepsilon) = \frac{g(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda)P(\omega_n)}{\sum_{\lambda',n',k'} g(y_\varepsilon|k', \omega_{n'}, \lambda')P(\omega_{n'}) + \delta} \quad (11)$$

The first term in the denominator is simply a weighted sum of Gaussian over all the speakers' multicondition GMMs; δ is a small positive number used to accommodate the noisy y_ε without matching subbands in the training data. Using this posterior probability, the frame likelihood (8) based on the full set of subband features can be written as

$$p(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda) \propto \sum_{n=0}^{N} P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y_\varepsilon) \quad (12)$$

Equation (12) is in a form suitable for feature selection. An optimal estimate of the matched feature set $y_\varepsilon(\omega_n)$, for each training noise condition ω_n, can be obtained by maximizing the corresponding posterior $P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y)$ over all possible sets $y \subseteq y_\varepsilon$. Denote by $\hat{y}_\varepsilon(\omega_n)$ such an estimate, $\hat{y}_\varepsilon(\omega_n) = \arg\max_{y \subseteq y_\varepsilon} P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|y)$. Then, the optimal frame likelihood, based on the optimal feature sets, can be expressed as

$$p(y_\varepsilon|k_{\mathbf{x},i_\varepsilon}, \lambda) \propto \sum_{n=0}^{N} P(k_{\mathbf{x},i_\varepsilon}, \omega_n, \lambda|\hat{y}_\varepsilon(\omega_n)) \quad (13)$$

Equation (13) combines multicondition training and optimal feature selection, to offer improved robustness to noise variations outside the training conditions. This frame likelihood is used to calculate segment likelihood (7) in the new LMS system for searching longest matching speech segments given noisy speech. A similar approach, termed universal compensation (UC), is proposed in [9] for a GMM framework. The above shows the extension of the UC technique into the new LMS framework. In the following experiments, we will compare the LMS approach and the GMM approach both equipped with UC-based noise compensation. Therefore, any differences in recognition accuracy between the two systems would be due to the modeling of the long-range temporal dynamics of speech in the LMS framework.

4. EXPERIMENTAL EVALUATION

The NIST SRE 2002 database, for the task of one speaker detection, was used in the experiments. The database contains cellular phone conversational speech data. The training set consists of 330 speakers (139 male, 191 female) with an average utterance length of about two minutes per speaker. In our experiments, we modeled each speaker using a GMM G_λ with 128 mixtures. For noise compensation, each Gaussian component in the clean GMM was expanded to a multicondition Gaussian set, modeling variable noise conditions (see Section 3). To simulate the unknown test noise, we corrupted the clean training data at 42 different noise conditions. These include low-pass filtered white noise with a bandwidth of 0.5, 1, 1.5, 2, 2.5 and 3 kHz, respectively, plus white noise without filtering; each noise type was added at six different SNR levels: 10, 12, 14, 16, 18 and 20 dB. These 42 simulated noise conditions, plus the clean condition, form a 43-condition GMM for each speaker. The GMM-based UC system [9] takes this multicondition GMM, combined with optimal subband selection at the decoding. The new LMS approach moves further, by creating a sentence model (6), defined on the multicondition GMM, to model the speaker; this training sentence model, combined with segment likelihood (7) and optimal subband selection (i.e., (13)), is used to identify the longest matching segments and perform recognition. Therefore, the two systems (noted as GMM+UC and LMS+UC) differ in the modeling of speech: LMS+UC models the speech temporal dynamics through the search for the longest matching segments, whereas GMM+UC assumes inter-frame independence.

There were a total of 3570 test sentences (1442 male, 2128 female) with variable durations from 15 to 45 s. Noisy test sentences were created by adding four different types of realistic noise at a SNR of 10 and 15 dB, respectively. The four noises were: an engine noise (taken from NoiseX92), a polyphonic musical ring, a pop song with mixed music with voice of a male singer, and voice of a female as crosstalk. These test noises each has a spectral structure significantly different from the multicondition training noises. The experimental results, thus, would demonstrate the ability of the proposed combination of multicondition model training and missing-feature theory over noise conditions unseen in training. While the engine noise exhibited some characteristics

Table 1: *Equal error rates (%) comparing baseline GMM, GMM+UC and the new LMS+UC approach.*

Noise type	SNR (dB)	GMM	GMM+UC	LMS+UC
Engine	10	35.18	28.57	23.07
	15	26.58	21.99	18.43
Musical ring	10	28.49	23.30	18.04
	15	21.87	19.12	15.55
Pop song	10	22.14	20.21	17.64
	15	18.18	17.70	15.52
Crosstalk	10	21.11	18.12	17.01
	15	18.15	16.91	15.70
Clean		14.90	16.44	14.05

of slow variation, the other two types of noise, song and musical ring, were highly nonstationary. The speech was divided into frames of 20 ms with a frame period of 10 ms. Each frame was modeled using 12 decorrelated log filterbank outputs uniformly divided into six subbands, with the addition of the corresponding first-order derivatives. We compared three recognition systems: 1) a baseline GMM system trained using clean data alone, 2) a GMM system with the UC method for noise compensation (GMM+UC) [9], and 3) the new LMS system with the UC method for noise compensation (LMS+UC). The comparison demonstrates that modeling the *noise* with the UC method (i.e., from GMM to GMM+UC) improves the recognition accuracy, and additionally, modeling the temporal dynamics of *speech* with the LMS method (i.e., from GMM+UC to LMS+UC) further advances the recognition accuracy.

Fig. 1– 5 present the DET curves comparing the three systems under each type of the test noises (including clean speech condition), as a function of the SNR. Table 1 summarizes the corresponding equal error rates (EER). As indicated in Fig. 1– 3 and Table 1, the GMM system with UC based noise compensation (GMM+UC) offered improved recognition accuracy over the baseline GMM in all the noise conditions. The new LMS+UC system, which combines modeling speech temporal dynamics and UC based noise compensation, further boosted the recognition accuracy from the GMM+UC system in all the noise conditions. The improvements by the LMS+UC system are quite significant in some noise conditions. For example, for both the musical ring and pop song noises, LMS+UC at SNR = 10 dB even outperformed GMM+UC at a higher SNR=15 dB (EER = 18.0% vs. 19.1%, and 17.6% vs. 17.7%, respectively). For the engine noise at SNR = 10 dB, LMS+UC reduced the EER of the baseline GMM by over 34% relatively. This is almost twice the reduction by the GMM+UC system ($\sim$18% relative). In the case of crosstalk noise which is even more difficult compared to other test noise conditions,the proposed system achieved over 6% and 7% relative improvement over GMM+UC for 10 and 15 dB SNR conditions respectively. As pointed out, all the improvements are due to the capture of long-range temporal dynamics of speech in the LMS+UC system. Finally, for clean speech test, the new LMS+UC outperformed the matched-condition baseline GMM. Further experiments using the clean trained LMS achieved an EER of 12.4% for clean data test. This is slightly better than our previous clean test LMS result based on fullband MFCC features [15].

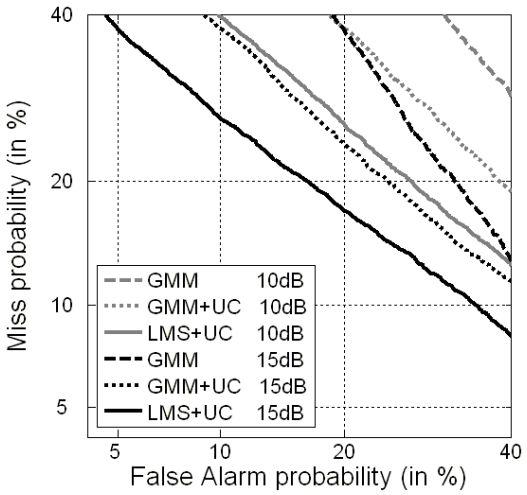

Figure 1: *DET curves for the engine noise, comparing baseline GMM, GMM+UC and the new LMS+UC, as a function of SNR (dB).*

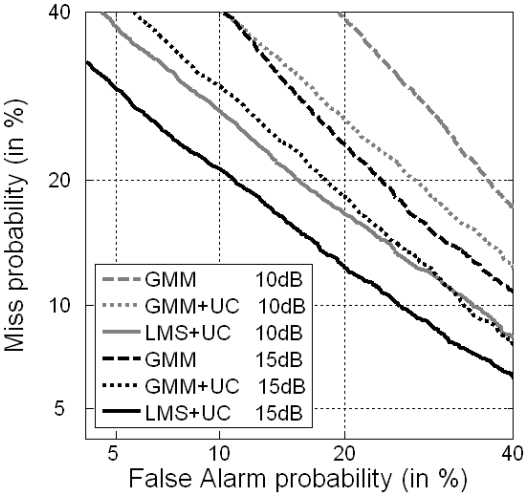

Figure 2: *DET curves for the musical ring noise.*

In this paper we have demonstrated that for speaker recognition under additive noise conditions, the LMS approach combined with the UC technique is superior to the GMM approach equipped with the same UC technique. This proves the importance of long term temporal dynamics for speaker recognition over short term features under noisy conditions. We have shown in [15] that MAP adaptation can be incorporated into the LMS speaker model; this has outperformed the conventional baseline GMM-UBM system and obtained an EER among the best for a single system for the NIST SRE 2002 task, without additional noise corruption. Hence there is a scope for further improving the noise robustness by incorporating maximum a posteriori (MAP) adaptation into the LMS+UC system described in this paper. This topic is under investigation.

5. CONCLUSION

We have presented an approach for robust speaker recognition in unknown/unpredictable noise environments. The new

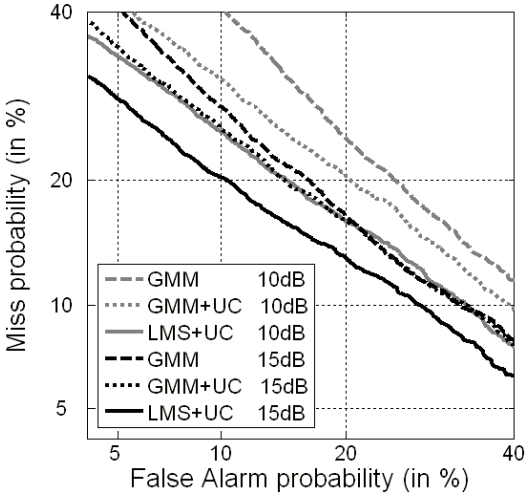

Figure 3: *DET curves for the pop song noise.*

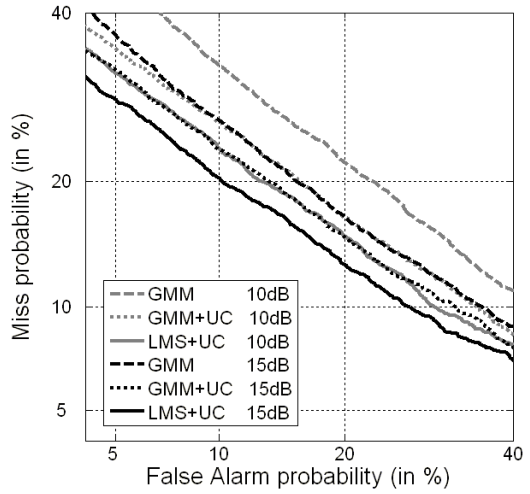

Figure 4: *DET curves for the crosstalk noise.*

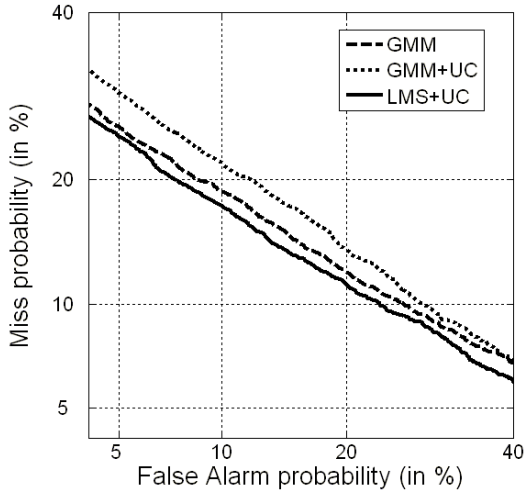

Figure 5: *DET curves for clean speech test comparing baseline GMM, GMM+UC and the new LMS+UC.*

method maximizes the extraction of the temporal dynamics of speech for speech and noise distinction. Experiments were conducted for speaker recognition in various noise conditions including fast-varying song and music. The new approach demonstrated significantly improved performance over conventional noise-robust techniques.

REFERENCES

[1] J. Ortega-Garcia and L. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," ICSLP'96, pp. 929932.

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

[3] Kwon, C. H., *et al.*, "Performance improvement of text-independent speaker verification systems based on histogram enhancement in noisy environments ", Interspeech'2008, pp. 1901-1904.

[4] Y. Shao, and D.L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis", ICASSP'2008, pp. 1589-1592.

[5] M. Wolfel, *et al.*, "Speaker identification using warped MVDR cepstral features," Interspeech'2009, pp. 912-915.

[6] L. Wang, K. Minami, K. Yamamoto, S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," ICASSP'2010, pp. 4502-4505.

[7] L. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," ICASSP'2001, pp. 457-460.

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

[9] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," IEEE Trans. Speech Audio Process., vol. 15, pp. 1711-1723, 2007.

[10] D. E. Sturim, *et al.*, "Speaker verification using text-constrained Gaussian mixture models," ICASSP'2002, pp. 667-680.

[11] H. Aronowitz, D. Burshtein, and A. Amir, "Text independent speaker recognition using speaker dependent word spotting," ICSLP'2004, pp. 1789-1792.

[12] Y. Tsao, *et al.*, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," ICASSP'2010.

[13] A. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey, "Modeling prosodic dynamics for speaker recognition," ICASSP'2003, pp. 788-791.

[14] W. D. Andrews, *et al.*, "Gender-dependent phonetic refraction for speaker recognition," ICASSP'2002, pp. 149-152.

[15] A. Jafari, R. Srinivasan, D. Crookes, and J. Ming, "A longest matching segment approach for text-independent speaker recognition," Interspeech'2010.