A Longest Matching Segment Approach with Baysian Adaptation - Application to 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
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 extract long-range temporal dynamics of speech for text-independent speaker recognition. We aim to maximize the noise immunity arising from the distinct temporal dynamics of speech. The new approach achieves this by identifying the *longest* matching segments between the training data and test data for recognition. Additionally, the new approach combines Bayesian adaptation, multicondition training and missing-feature theory to further advance the ability to model noisy speech. 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 improved performance over conventional noise-robust techniques.
Index Terms: Temporal dynamics, noise robustness, speech segmentation, speaker recognition.

1. Introduction

In present speaker recognition systems, methods to reduce the influence of background noise include one or more combinations of the following: 1) speech enhancement [1], 2) robust acoustic features [2, 3], and 3) noise compensation (e.g., parallel model combination, multicondition model training, and missing-feature theory) [4, 5, 6]. 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 an open problem. People have studied text-constrained speaker recognition, based on common subword or word units between the training and test data [7, 8]. Alternatively, recognition has been based on acoustic segments identified either by phonetic similarity or by minimum distance [9]. Other methods include the use of prosodic features [10] and phonetic refraction, expressed as phone *n*-gram counts [11].

In this paper, we present an approach to model long-range temporal dynamics of speech with the aim of improving the noise robustness. We achieve this by identifying and comparing the *longest* matching segments between the training data and the test data. Longer speech segments as whole units contain more distinct temporal dynamics, and hence can be distinguished more accurately from noise than shorter speech segments. Therefore recognition based on the longest matching segments increases the noise immunity, and hence reduces the requirement for information about the noise. For convenience, we call the new approach the longest matching segment (LMS) approach. This work is an extension of our previous work [12], which was for clean speech, to noisy speech. The extensions include two main parts: 1) a new LMS framework derived from Bayesian adaptation, which allows the construction of robust LMS with limited training data, and 2) integrating multicondition training and missing-feature theory into the adapted LMS framework, to further improve the noise robustness assuming minimal information about the noise.

2. Longest Matching Segment Framework

We use a data-driven approach to model long-range temporal dynamics in the training data, and a statistical approach to identify the longest matching segments between the training data and test data. To model a speaker, first we estimate a GMM using all the training data from the speaker, as in normal GMM-based recognition systems. Let $G_\lambda = \{g(x|k,\lambda), w(k|\lambda) : k = 1,2,...,K_\lambda\}$ denote the GMM for speaker λ, where $g(x|k,\lambda)$ is the k'th Gaussian component and $w(k|\lambda)$ is the weight. Next, based on G_λ, we build a model for each training utterance from the speaker. This model represents the *full* temporal dynamics in the training utterance, and facilitates the identification of the corresponding temporal dynamics in test utterances by using a statistical approach. Let $\mathbf{x} = \{x_i : i = 1,2,...,I_\mathbf{x}\}$ be a training utterance from speaker λ, with $I_\mathbf{x}$ frames. We represent $\mathbf{x}$ by using a corresponding time sequence consisting of the Gaussian components that produce the maximum likelihood for $\mathbf{x}$. This time sequence can be expressed as

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

where $(k_{\mathbf{x},i}, \lambda)$ is the index of the Gaussian component $g(x|k_{\mathbf{x},i}, \lambda)$ in G_λ producing the maximum likelihood for frame x_i in the training utterance $\mathbf{x}$ from speaker λ. This Gaussian sequence shares characteristics with a traditional template but uses a statistical approach (i.e., GMM) to capture the full temporal dynamics in $\mathbf{x}$. In training, we create one such model

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

for each training utterance; all the training utterance models $(\mathbf{k}_\mathbf{x}, \lambda)$ of a speaker together form the model for the speaker, to be used for identifying the matching segments in the test utterances for recognition.

Let $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$ represent a test utterance with T frames, and $\mathbf{y}_{t:\tau} = \{y_\epsilon : \epsilon = t, t+1, ..., \tau\}$ be a test segment in $\mathbf{y}$ starting at time t and consisting of consecutive frames from t to τ. In a similar notation, let $(\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \{(k_{\mathbf{x},i}, \lambda) : i = u, u+1, ..., v\}$ represent a training segment, taken from the model $(\mathbf{k}_\mathbf{x}, \lambda)$ and modeling the consecutive frames from u to v in the training utterance $\mathbf{x}$ of speaker λ. The recognition score for speaker λ given $\mathbf{y}$ can be calculated as follows:

$$\Gamma(\lambda; \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} \max_{\tau} \max_{\mathbf{k}_{\mathbf{x},u:v} \in \lambda} \log P(\mathbf{k}_{\mathbf{x},u:v}, \lambda | \mathbf{y}_{t:\tau}) \quad (2)$$

where $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda | \mathbf{y}_{t:\tau})$ is the probability that the training segment $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ matches the test segment $\mathbf{y}_{t:\tau}$. At each test frame at t, the longest matching segments are obtained by jointly maximizing the probability over all training segments for each fixed-length test segment and over all possible test segment lengths (i.e., τ). The scores of the longest matching segments found at all the test frames are combined for recognition.

Using Bayes' formula, the probability of the matching training segment given a test segment can be written as

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

where we assume an equal prior probability for each training segment; $p(\mathbf{y}_{t:\tau} | \mathbf{k}_{\mathbf{x},u:v}, \lambda)$ is the likelihood that the test segment $\mathbf{y}_{t:\tau}$ is matched by the training segment $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$; $p(\mathbf{y}_{t:\tau} | \phi)$ represents the likelihood that $\mathbf{y}_{t:\tau}$ is matched by a segment not found in the given training data. This likelihood of unseen test segments can be modeled by a GMM trained using the training data from all the speakers. Assuming independence between the frames within a segment, the segmental likelihood $p(\mathbf{y}_{t:\tau} | \mathbf{k}_{\mathbf{x},u:v}, \lambda)$ can be expressed as

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

where i_ϵ is the most-likely time warping path between the two segments. It is shown in [12] that (3) favors the continuity of the segment match, by giving a larger probability of match for longer matching training and test segments. Since long matching speech segments can be identified more accurately in noise than short matching speech segments, the above LMS algorithm, which performs recognition based on the longest matching segments identified, has the potential to improve the noise robustness.

3. Improving LMS Robustness

In this paper, we extend the above LMS framework for noise-robust speaker recognition, assuming minimal prior knowledge about the noise. We achieve this by introducing Bayesian adaptation into the LMS framework, and by integrating multicondition training and missing-feature theory in the adapted LMS framework to further advance its ability to model noisy speech.

3.1. Adapted LMS

The GMM G_λ, and the associated training utterance models $(\mathbf{k}_\mathbf{x}, \lambda)$ [i.e., (1)], for each speaker can be derived from a universal background model (UBM) using Bayesian adaptation [13]. This is a three step process. First, we use the speaker's training data to update the UBM using the EM algorithm, obtaining a reestimation-based GMM for the speaker. Second, on the reestimation-based GMM, we create the utterance model (1) for each training utterance of the speaker, where each index $(k_{\mathbf{x},i}, \lambda)$ defines an alignment between a training frame and a Gaussian component in the reestimation-based GMM. We identify the alignments and hence the training utterance models at this stage to make sure that only those Gaussian components receiving training frames from the speaker are used to model the speaker. Finally, the reestimation-based Gaussian components are combined with the UBM Gaussian components to form the adapted GMM for the speaker, to improve the robustness of those Gaussian components without sufficient training data.

3.2. Adapted LMS with multicondition training

We simulate the test noise by adding variable forms of noise to the clean training utterances. Let ω_n, $n = 0, 1, ..., N$, represent $N+1$ training noise conditions, with ω_0 representing the clean condition. Multicondition training can be introduced into the adapted LMS efficiently, based on the UBM and the training utterance models generated from the clean training data, described above. First, we obtain a UBM for each training noise condition using the corresponding noisy training data, by copying the alignments between the training frames and Gaussian components obtained for the clean UBM. Hence all the UBMs have the same frame to Gaussian component alignments. In a similar way, for each speaker, we can obtain an effective reestimation-based GMM for each training noise condition by using the corresponding noisy training data from the speaker; this can be accomplished efficiently by using the training frame to Gaussian component alignments: $x_i \rightarrow (k_{\mathbf{x},i}, \lambda)$, specified in the speaker's training utterance models (1). Therefore, the same Gaussian components at different training conditions are estimated using the same frames corrupted at the respective conditions. Finally, an adapted GMM is obtained for each training noise condition by combining the UBM and reestimation-based GMM obtained for that condition. Thus, we have expanded the clean adapted GMM G_λ for each speaker λ into a multicondition adapted GMM, in which each clean Gaussian component $g(x|k, \lambda)$ is expanded into a set of $N+1$ Gaussian components $g(x|k, \omega_n, \lambda)$, $n = 0, 1, ..., N$, where $g(x|k, \omega_n, \lambda)$ is adapted using the training frames corresponding to $g(x|k, \lambda)$ but corrupted at the noise condition ω_n. Based on this multicondition adapted GMM, we can obtain a new utterance model for each training utterance $\mathbf{x}$, which has the same form as (1), but each frame index $(k_{\mathbf{x},}, \lambda)$ now defines a *multicondition* frame likelihood $p(x|k_{\mathbf{x},}, \lambda)$, which can be expressed as

$$p(x|k_{\mathbf{x},i}, \lambda) = \sum_{n=0}^{N} g(x|k_{\mathbf{x},i}, \omega_n, \lambda) P(\omega_n) \quad (5)$$

Eq. (5) takes into consideration the variable frame corruption ω_0 through ω_N, with $P(\omega_n)$ being the prior of the corruption (assumed to be a uniform distribution in our experiments). In the new LMS system, this multicondition likelihood, refined by optimal feature selection to be discussed next, is used to replace the single-condition Gaussian likelihoods in (4) to calculate the segment likelihoods, to reduce noise-caused mismatches.

3.3. Optimal feature selection

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 $y_\epsilon = (y_{\epsilon,1}, y_{\epsilon,2}, ..., y_{\epsilon,F})$, where $y_{\epsilon,f}$ 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, $y_\epsilon(\omega_n)$, includes the subband features that are matched by the training noise condition ω_n; the other subset, the complement $\tilde{y}_\epsilon(\omega_n)$, 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 (5) based on the matched feature sets (i.e., the missing-feature theory). An optimal estimate for the matched feature set $y_\epsilon(\omega_n)$ in frame y_ϵ can be obtained by expressing $g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)P(\omega_n)$ in (5) through a posterior probability and maximizing the posterior probability over all possible feature subsets. The posterior probability proportional to $g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)P(\omega_n)$ can be obtained as follows

$$\begin{aligned} g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)P(\omega_n) &= \frac{g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)P(\omega_n)}{p(y_\epsilon)} p(y_\epsilon) \\ &= P(k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda|y_\epsilon)p(y_\epsilon) \end{aligned} \quad (6)$$

The last term $p(y_\epsilon)$ is not a function of the matching training frame and hence can be ignored; $P(k_{\mathbf{x},i}, \omega_n, \lambda|y_\epsilon)$ is the posterior probability of the training frame $(k_{\mathbf{x},i}, \omega_n, \lambda)$ that matches the test frame y_ϵ, which can be expressed as

$$P(k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda|y_\epsilon) = \frac{g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)P(\omega_n)}{\sum_{\lambda',k',n'} g(y_\epsilon|k', \omega_{n'}, \lambda')P(\omega_{n'}) + \delta} \quad (7)$$

where δ is a small positive number used to account for the noisy y_ϵ without matching subbands in the model data. An optimal estimate of the matched test subset $y_\epsilon(\omega_n)$ for the training frame $(k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda)$ can be obtained by maximizing $P(k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda|y)$ over all possible sets $y \subseteq y_\epsilon$. The frame likelihood (5) with the feature optimization can be written as

$$p(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \lambda) \propto \sum_{n=0}^{N} \max_{y \subseteq y_\epsilon} P(k_{\mathbf{x},i_\epsilon}, \omega_n, \lambda|y) \quad (8)$$

The above model (8) combines multicondition training and missing-feature theory to offer robustness to variable noise within and outside the training conditions. A similar approach, termed universal compensation (UC), was proposed in [6] for a GMM framework. The above shows the extension of the UC technique into the adapted LMS framework.

4. Experimental studies

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 trained a UBM with 1024 diagonal-covariance Gausssian components using the clean data, and obtained an adapted GMM, along with the training utterance models, for each speaker using the algorithm described in Section 3.1. Further, to simulate unknown test noise, we corrupted the clean training utterances by adding low-pass filtered white noise at different SNR levels: 12, 14, 16, 18 dB. These four noisy training conditions, plus the original clean training condition, result in five training conditions. This was followed by the creation of a multicondition adapted GMM and training utterance models for each speaker, using the algorithm described in Section 3.2. 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 three different types of realistic noise at a signal-to-noise ratio (SNR) of 10 and 15 dB, respectively. The three noises were: an engine noise, a musical ring, and a pop song with mixed music and a male singer. 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 energies uniformly divided into six subbands, with the addition of the first-order derivatives. We compared three speaker recognition systems: 1) a baseline GMM-UBM trained using clean data alone (called GMM), 2) a GMM-UBM with UC-based noise compensation (called GMM+UC), and 3) the new adapted LMS system with UC-based noise compensation (called LMS+UC). The comparison demonstrates that modeling the *noise* with the UC approach 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.

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

Noise	SNR (dB)	GMM	GMM+UC	LMS+UC
Engine	10	34.79	26.38	20.15
	15	26.27	20.53	16.81
Musical ring	10	27.67	21.44	17.72
	15	20.50	17.87	15.53
Pop song	10	20.56	21.07	18.42
	15	14.50	18.02	15.86

Table 2: *EER (%) for clean test data (using MFCC features).*

System	GMM	LMS
EER%	9.95	9.36

Table 3: EER (%) for clean-data trained GMM and LMS, and noise compensated GMM+UC and LMS + UC, for musical ring noise.

SNR (dB)	GMM	LMS	GMM+UC	LMS+UC
10	27.67	24.25	21.44	17.72
15	20.50	17.59	17.87	15.53

Fig. 1– 3 present the DET curves comparing the three systems under each type of test noise, as a function of the SNR. Table 1 summarizes the corresponding equal error rates (EER). As can be seen, the GMM+UC system offered improved recognition accuracy over the baseline GMM in almost all the noise conditions, except for the speech-like 'pop song' noise. The new LMS+UC system further boosted the recognition accuracy from the GMM+UC system, and only showed slightly worse accuracy for the pop song noise at the higher SNR case, than the baseline GMM. This case where no improvement is found for the pop song noise indicates the difficulty of modeling speech-like noise for accurate speaker recognition. The improvements by the LMS+UC system over the baseline and GMM+UC for non-speech noise are significant in all non-speech noise condi-

tions. For example, for both the engine and musical ring noises, LMS+UC at SNR = 10 dB even outperformed GMM+UC at a higher SNR=15 dB (EER = 20.1% vs. 20.5%, and 17.7% vs. 17.9%, respectively). For the engine noise at SNR = 10 dB, LMS+UC reduced the EER of the baseline GMM by over 42% relatively. This is almost twice the reduction by the GMM+UC system ($\sim$24% relatively). As pointed out, all the improvements over the GMM+UC 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 outperformed the baseline GMM, as shown in Table 2. To the best of our knowledge, the EER of 9.36% obtained by the new LMS system for the clean test data represents one of the best results obtained on the database by a stand-alone system.

Further comparisons are made between the baseline GMM and the new LMS, with and without noise compensation, for the musical ring noise with SNR = 10 and 15 dB, shown in Table 3. Two observations can be drawn from the table: 1) capturing long-range temporal dynamics of speech improved the noise robustness (i.e., LMS vs. GMM, and LMS+UC vs. GMM+UC), and 2) combining multicondition training and missing-feature theory to model noise further improved the robustness (i.e., LMS+UC vs. LMS).

5. Conclusions

A new speaker recognition approach based on detecting the longest matching segments between training and test utterances was described. As long matching speech segments can be detected more accurately from noise, the new approach should offer improved noise robustness compared to conventional GMM-based systems. This was demonstrated in this paper. The paper described the integration of Bayesian adaptation, multicondition training and missing-feature theory into the new framework to further improve the noise robustness. Experiments were conducted on the NIST SRE 2002 database with the addition of various types of nonstationary noise corruption. The results indicate that the proposed system outperformed the baseline GMM-UBM system based on the same noise compensation techniques. The improvement is due to the enhanced ability of the new method to capture long-range temporal dynamics of speech. We also showed that it can be more difficult to handle speech-like noise than non-speech noise with traditional noise compensation methods.

6. References

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

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

[3] L. Wang, *et al.*, "Speaker identification by combining MFCC and phase information in noisy environments," ICASSP'2010.

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

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

[6] 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.

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

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

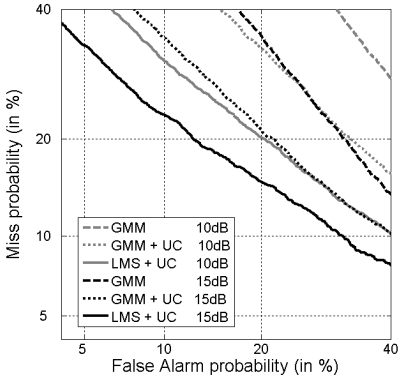

Figure 1: *DET curves for the engine noise, comparing baseline GMM, GMM+UC and the new LMS+UC, as a function of SNR (dB).*

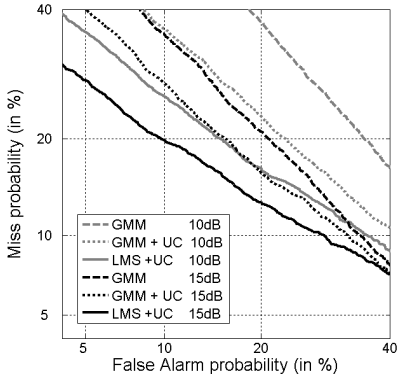

Figure 2: *DET curves for the musical ring noise.*

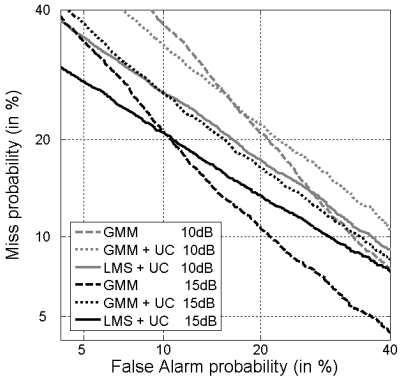

Figure 3: *DET curves for the pop song noise.*

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

[10] A. Adami, *et al.*, "Modeling prosodic dynamics for speaker recognition," ICASSP'2003, pp. 788-791.

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

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

[13] D.A. Reynolds, *et al.*, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process., vol. 10, pp. 19-41, 2000.