On Speech Features Fusion, α-Integration Gaussian Modeling and Multi-Style Training for Noise Robust Speaker Classification

A. Venturini, L. Zão, *Member, IEEE*, and R. Coelho, *Member, IEEE*

***Abstract*—This paper investigates the fusion of Mel-frequency cepstral coefficients (MFCC) and statistical pH features to improve the performance of speaker verification (SV) in non-stationary noise conditions. The α-integrated Gaussian Mixture Model (α-GMM) classifier is adopted for speaker modeling. Two different approaches are applied to reduce the effects of noise corruption in the SV task: speech enhancement and multi-style training (MT). The spectral subtraction with minimum statistics (MS/SS) and the optimally-modified log-spectral amplitude with improved minima controlled recursive averaging (IMCRA/OMLSA) are examined for the speech enhancement procedure. The MT techniques are based on colored (Colored-MT), white (White-MT) and narrow-band (Narrow-MT) noises. Six real non-stationary noises, collected from different acoustic sources, are used to corrupt the TIMIT speech database in four different signal-to-noise ratios (SNR). The index of non-stationarity (INS) is chosen for the stationarity tests of the acoustic noises. Complementary SV experiments are conducted in realistic noisy conditions using the MIT database. The results show that the best SV accuracy was obtained with the MFCC + pH features fusion, the MS/SS and the Colored-MT.**

***Index Terms*—Features fusion, Hurst exponent, multi-style training, non-stationary acoustic noise, speaker verification, speech enhancement, α-GMM.**

I. Introduction

SPEAKER verification or authentication is an interesting solution for applications with security concerns, such as access control, data security and forensic investigations [1] [2]. Recently, it has become more evident the need for security solutions for portable devices, such as laptops and smartphones. It implies that SV systems must keep good performance at different conditions, even in adverse noisy environments.

The conventional speaker verification framework, which adopts the GMM [3] classifier and the universal background model (UBM) [4], is generally composed of a training and a test phase [5]. During training, the system extracts some sets of speech features and generates the UBM and the speakers models. In the test phase, the speech features are obtained from the test utterances and compared to the speakers models. The main goal of the SV task is to decide whether to accept or reject a claimed identity. SV systems based on MFCC [6] features and GMM-UBM generally achieve high recognition accuracies for clean speech [7]. However, their performance can be severely degraded when the speech signals are corrupted by acoustic noises [8].

Recently, expanded versions of the GMM-UBM SV systems have been proposed in the literature. In [9], support vector machines (SVM) were applied in the GMM supervector space using Nuisance Attribute Projection (NAP) [10] to compensate for the channels effect. Other state-of-the-art SV systems that deal with channel compensation are the eigenvoice [11] and the joint factor analysis (JFA) [12]. In [13], the factor analysis was also used to define a new set of features called *i-vectors*. Currently, most of the SV systems use the *i-vectors* together with the probabilistic linear discriminant analysis (PLDA) [14] to produce good performance in clean conditions. However, the analysis of such systems in noisy scenarios is rarely found in the literature.

One of the most interesting solutions for robust speaker recognition submitted to real acoustic noises is the multi-style training (MT) [15], [16], [8], [17]. The MT was originally proposed to improve the speech recognition in noise [15]. The idea is to reduce the training and test mismatch by corrupting the utterances available for training. In [8], assuming no information about the noise sources, the authors proposed the use of artificial white and narrow-band noises to corrupt the training speech with different SNR values and improve speaker recognition performance. In [17], artificial noises with colored spectra and a single SNR value were adopted to corrupt the training utterances and provide robustness to speaker identification. This colored-noise-based multi-style training was also applied in [18] for the speaker verification task.

In this paper, the MT based on colored (Colored-MT) [17], white (White-MT) and narrow-band (Narrow-MT) [8] noises are applied with the α-GMM [19] classifier to improve the noise robustness of speaker verification. The conventional GMM is considered as a particular case of the α-GMM ($\alpha = -1$). Besides the MFCC and their corresponding velocity (Δ) and acceleration ($\Delta\Delta$) coefficients, the pH [20] statistical feature is also used to compose the speech feature vectors. In [20], the authors showed that the fusion of the MFCC and pH features improves

Manuscript received November 04, 2013; revised March 24, 2014; accepted August 26, 2014. Date of publication September 08, 2014; date of current version September 20, 2014. This work was supported in part by the CNPq under research grants 472461/2009-5 and 304254/2012-6. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rodrigo S. Guido.

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering (IME), Rio de Janeiro RJ, 22290-270, Brazil (e-mail: coelho@ime.eb.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2355821

the performance of single MFCC-based speaker verification and identification tasks, when submitted to telephone channel distortion. One of the main goals of this paper is to investigate the contribution of the pH feature to reduce the equal error rate (EER) of the speaker verification systems in non-stationary acoustic noisy environments. For this purpose, SV experiments are conducted with speech utterances, collected from the TIMIT database [21], corrupted by six acoustic noises (Babble, Engine, Factory, Machine Gun, Military Vehicle, and Ringtone) with different values of the index of non-stationarity (INS) [22], considering SNR of 5, 10, 15 and 20 dB. Moreover, the pH feature is also evaluated in SV experiments conducted in realistic noisy conditions using the MIT Mobile Device Speaker Verification Corpus [23].

Another solution adopted in this work to reduce the noise effects in SV is to enhance the speech signals. A common approach used in speech enhancement involves a voice activity detector (VAD) and the short-time spectrum analysis (STSA) to estimate the noise spectral components from segments with no speech presence. If the noise is stationary, these components can then be suppressed from the entire speech by spectral subtraction (SS) [24], [25]. However, when the stationarity of the noise is not assured, the estimated noise spectrum must be updated even during voice activity [26]. The minimum statistics (MS) [27] and the improved minima controlled recursive averaging (IMCRA) [28] techniques were proposed to follow such requirement. Speech enhancement techniques have been evaluated for speaker verification [29] and identification [30] considering stationary background noises. In this work, two speech enhancement procedures are used in the SV experiments to enhance the speech signals in non-stationary noise environments. The first one adopts the spectral subtraction combined with the MS technique. In the second method, the IMCRA is followed by the optimally-modified log-spectral amplitude (OMLSA) [31] speech estimator. Both MS/SS and IMCRA/OMLSA are evaluated as pre-processing steps for the speaker verification system.

The remainder of this work is organized as follows. Section II provides a general description of a speaker verification system. It includes the MFCC and pH speech features and the α-GMM classifier. The same Section presents the colored-noise-based multi-style training. The basic concepts of the MS/SS and IMCRA/OMLSA speech enhancement techniques are presented in Section III. Section IV describes the speaker verification experiments evaluated in different noisy environments with TIMIT database. The experiments are conducted with and without the Colored-MT and speech enhancement. In Section V, these same techniques are evaluated with the MIT database. Finally, Section VI concludes this work.

II. Speaker Verification

The main issue of this paper is to evaluate the contribution of different techniques in the speaker verification task. It includes the MFCC and pH features fusion, the α-GMM classifier as an alternative to the regular GMM, and the use of speech enhancement algorithms with multi-style training. The SV experiments are conducted in a conventional GMM-UBM system without using any other score normalization or channel compensation approaches. The recent expansions of the GMM-based system, such as the JFA or the *i-vectors* with PLDA, are not used since they would require a highly time consuming training step for each of the SV experiments conducted in this work.

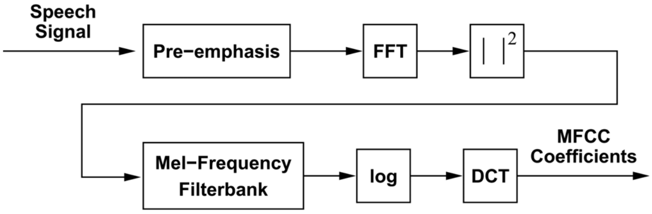


Fig. 1. Representation of the MFCC extraction.

This Section describes the verification function adopted in this work, including the speech features used for the speakers representation, the α-GMM and also the multi-style training proposed for the noise robust speaker classification. The main goal of the verification task is to decide whether the observed speech segment belongs or not to the claimed speaker. Given the feature matrix $\mathbf{X}$ obtained from the speech signal Φ and a hypothesized speaker L, the SV decision corresponds to the hypothesis test between

$$\begin{cases} H_0: & \mathbf{X} \text{ belongs to } L, \\ H_1: & \mathbf{X} \text{ does not be long to } L, \end{cases} \tag{1}$$

The commonly used criteria to decide if Φ belongs to L is based on the likelihood ratio test,

$$\frac{p(\mathbf{X}|\lambda_L)}{p(\mathbf{X}|\lambda_{\text{UBM}})} \begin{cases} \geq \theta, & \text{accept } H_0, \\ < \theta, & \text{reject } H_0, \end{cases} \tag{2}$$

where λ_L represents the model of speaker L, λ_{UBM} is the universal background model (UBM) and θ is the decision threshold.

In (2), $p(\mathbf{X}|\lambda_L)$ is the probability density function (pdf) of $\mathbf{X}$ given it was spoken by the claimed speaker L. In the same way, $p(\mathbf{X}|\lambda_{\text{UBM}})$ is the pdf of $\mathbf{X}$ given that it is not from the claimed speaker, i.e., the speech segment belongs to an intruder. The UBM is a single model that is generally obtained from the speech of many speakers that are not enrolled to the system.

The choice of the decision threshold (θ) is a tradeoff between the false rejection (FR) and false acceptance (FA) errors. These probabilities are usually evaluated by detection error tradeoff (DET) curves. In this work, the equal error rate is used to measure the performance of the SV system. The EER corresponds to the operating point where the probabilities of false rejection (P_{FR}) and false acceptance (P_{FA}) are equal. The NIST 2012 Speaker Recognition Evaluation Plan [32] defines the primary cost function (C_{primary}) as a combination of P_{FR} and P_{FA}. In this work, the minimum value of C_{primary} (min C_{primary}) is used as a complementary SV performance measure.

A. Speech Features

After the signal pre-processing (digitization, quantization and pre-emphasis), the speech features are extracted or estimated from short-time (20 ms - 30 ms) duration frames [33]. In this work, the speech feature matrices are formed by MFCC (with Δ and $\Delta\Delta$) and also by the MFCC + pH features fusion.

1) MFCC: The Mel-frequency cepstral coefficients [6] are the most widely used features for speaker verification. Fig. 1 depicts the schematic of the MFCC extraction.

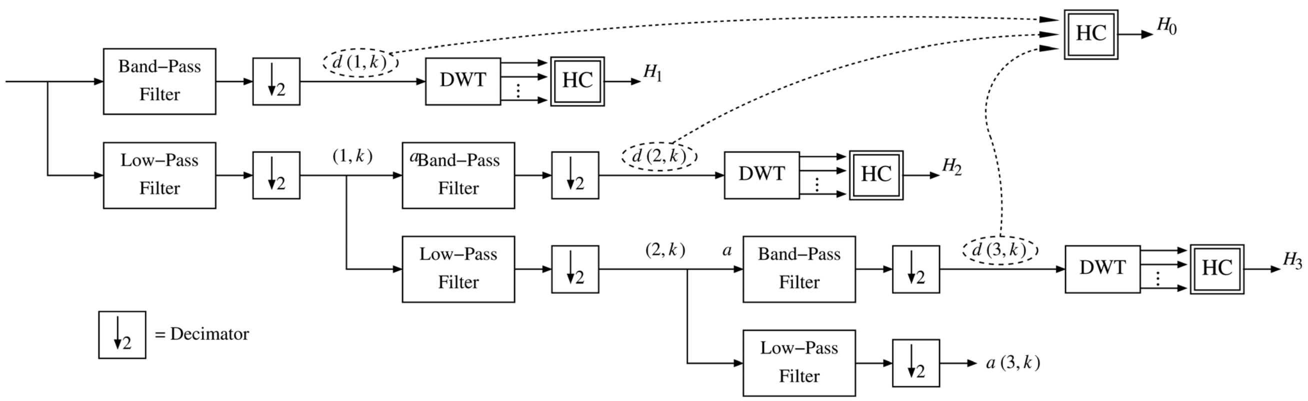


Fig. 2. An example of a pH vector estimation using the $M\_dim\_wavelets$ with 3 decomposition stages.

After the acquisition/pre-processing step, the fast Fourier transform (FFT) is calculated from each speech frame. The spectral envelope of the modulus of the FFT is then obtained using Mel-scaled bandpass filters. Frequencies in the Mel scale (f_{Mel}) are related to frequencies in the linear scale (f_{Hz}) as $f_{\mathrm{Mel}} = 1127 \cdot \log(1 + f_{\mathrm{Hz}}/700)$. The Mel scale is usually applied in speaker verification due to its good representation of the human auditory system. The MFCC are finally calculated by applying the discrete cosine transform (DCT) to the log-energy outputs of the filters in the Mel-frequency filterbank.

2) pH: The pH is a vocal time-frequency feature and was proposed and evaluated for speaker identification and verification systems [20]. It consists of a vector of Hurst ($0 \le H \le 1$) values, which expresses the time-dependence or scaling degree of the speech signal.

Let the speech signal be represented by a stochastic process $y(t)$, with the normalized autocorrelation coefficient function defined as

$$\rho(\mathcal{T}) = \frac{Cov\left[y(t), y(t+\mathcal{T})\right]}{Var\left[y(t)\right]}. \tag{3}$$

The Hurst is defined by the decaying rate of $\rho(\mathcal{T})$, whose asymptotic behavior is given by

$$\rho(\mathcal{T}) \sim H(2H-1)\mathcal{T}^{2(H-2)},\ \mathcal{T} \to \infty. \tag{4}$$

The Hurst exponent is related to the spectral characteristics of the speech signal. Within the whole range $]0, 1[$, the power spectral density (PSD) of $y(t)$, $S_y(f)$, can be shown to be proportional to f^{1-2H} when $f \to 0$. For $H = 1/2$, $S_y(f)$ is constant over the whole frequency spectrum (e.g., white noise), whereas low frequencies are prominent in the case where $H > 1/2$, and in particular when $H \to 1$ ($1/f$ or pink noise).

The wavelet-based multi-dimensional estimator (M-dim-wavelets) [20] was proposed as the pH feature extractor and is based on the method described in [34]. The estimation procedure is as follows:

- Wavelet decomposition: discrete wavelet transform (DWT) is applied to successively decompose a sequence of samples into approximation ($a(j,k)$) and detail ($d(j,k)$) coefficients, where j is the decomposition scale ($j = 1, 2, \ldots, J$) and k is the coefficient index of each scale.
- Hurst computation (HC): for each scale j, the variance $\sigma_j^2 = (1/n_j)\sum_k d(j,k)^2$ is evaluated from the detail coefficients, where n_j is the number of available coefficients for each scale j. In [34], it is shown that $E[\sigma_j^2] = \mathcal{C}_H \cdot j^{2H-1}$, where $\mathcal{C}_H$ is a constant. A weighted linear regression is then used to obtain the slope a of the plot of $y_j = \log_2(\sigma_j^2)$ versus j. The value of H is given by $H = (1+a)/2$.
- pH vector composition: the pH vector is composed of $(J+1)$ values of $H[H_0, H_1, \ldots, H_J]$. The H_0 component is computed from the decomposition of the entire speech signal. The other values $(H_1, H_2, \ldots, H_J)$ are obtained after re-applying the DWT decomposition to each of the J detail sequences. Fig. 2 shows an example of the pH estimation considering $J = 3$ decomposition stages, i. e., $[H_0, H_1, H_2, H_3]$.

In this work, the Daubechies wavelets filters [35] are applied for the DWT decomposition. The multi-resolution analysis [36] adopted in the DWT is a powerful theory that enables the detail and approximation coefficients to be easily computed by a simple discrete time convolution. It is important to note that the linear computational complexity of the pyramidal algorithm to obtain the DWT is $O(n)$ where n is the signal samples length, while the FFT (fast Fourier transform), used to obtain the MFCC, is $O(n\log(n))$.

B. α-GMM

In [19], the authors proposed the α-integration of Gaussian densities as an extension of the classical GMM for speakers modeling in a speaker identification task. The α-integration generalizes the linear combination adopted in the conventional GMM. The essential idea was to improve the identification performance by emulating an integration process similar to what occurs inside a human brain. In [19], it was shown that the α-GMM outperforms the conventional GMM in speaker identification task with speech transmitted through a fixed phone channel. In [18], the α-GMM was applied to a speaker verification task in noisy conditions, and it also achieved better results than the GMM.

Given a speaker model λ_L, composed of M Gaussian densities $b_i(\mathbf{x})$, $i = 1, \ldots, M$, the α-integration of the densities is defined as [19]

$$p(\mathbf{x}|\lambda_L) = \mathcal{C} f_\alpha^{-1}\left\{\sum_{i=1}^{M} \pi_i f_\alpha\left[b_i(\mathbf{x})\right]\right\}, \tag{5}$$

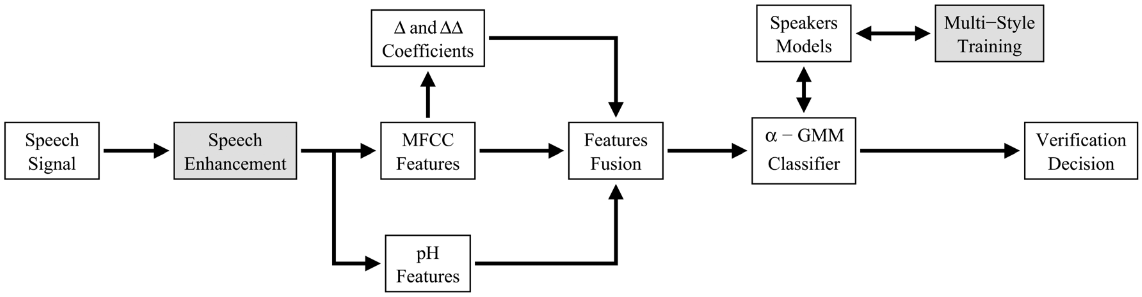


Fig. 3. Block diagram with the general work-flow adopted in this paper. The highlighted blocks indicate the two techniques used to improve the performance of the speaker verification in noise.

where π_i are non-negative mixture weights constrained to $\sum_{i=1}^{M} \pi_i = 1$, $\mathcal{C}$ is a normalization constant, and $f_\alpha(\cdot)$ is given by

$$f_\alpha(x) = \begin{cases} \left(\frac{2}{1-\alpha}\right) x^{(1-\alpha)/2}, & \alpha \neq 1, \\ \log(x), & \alpha = 1. \end{cases} \tag{6}$$

From (6), the inverse of $f_\alpha(\cdot)$ can be calculated by

$$f_\alpha^{-1}(y) = \begin{cases} \left(\frac{1-\alpha}{2} y\right)^{\frac{2}{1-\alpha}}, & \alpha \neq 1, \\ \exp(y), & \alpha = 1. \end{cases} \tag{7}$$

The α-GMM of speaker L can be rewritten as

$$p(\mathbf{x}|\lambda_L) = \mathcal{C}\left[\sum_{i=1}^{M} \pi_i b_i(\mathbf{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}. \tag{8}$$

Note that the α-integration of the Gaussian densities in (8) turns into a linear combination for $\alpha = -1$, which corresponds to the conventional GMM. By choosing values of α smaller than -1, the α-GMM classifier emphasizes the larger probability values, and de-emphasizes the smaller ones. The idea of this work is to use this property to compensate the training and test mismatch caused by environmental acoustic noises.

The α-GMM of a speaker L is completely parametrized by the mean vectors ($\vec{\mu}_i$), covariance matrices (K_i) and the weights of the Gaussian densities,

$$\lambda_L = \{\pi_i, \vec{\mu}_i, K_i | i = 1, \ldots, M\}. \tag{9}$$

Such parameters are estimated using the adapted expectation-maximization (EM) algorithm [37] as to maximize the likelihood function

$$p(\mathbf{X}|\lambda_L) = \prod_{t=1}^{Q} p(\mathbf{x}_t|\lambda_L), \tag{10}$$

where $\mathbf{x}_t$, $t = 1, \ldots, Q$, are the vectors extracted from the training speech segment Φ_L of speaker L, that compose the feature matrix $\mathbf{X}$. The likelihood expressed in (10) is also used in the likelihood ratio test in (2) for the decision criteria.

C. Multi-Style Training Based on Colored Noises

The multi-style training technique consists of artificially corrupting the training utterances to reduce the mismatch between the training and test phases, and thus, improve the performance of the speakers classification systems. In the colored-noise-based multi-style training (Colored-MT) [17], artificial noises are generated with Gaussian distribution and PSD characterized by the shape $S(f) \propto 1/f^\beta$, with $\beta \in [0, 2]$ [38].

For each speaker L, multiple copies of the clean training utterance Φ_L are corrupted by the artificial colored noises, resulting in multicondition data sets Φ_L^l ($l = 1, \ldots, m$). Following the procedure addressed in Section II-B, m α-GMM (λ_L^l) for speaker L are obtained from the corrupted data sets Φ_L^l. In analogy to (9), each of these models are parametrized by

$$\lambda_L^l = \{\pi_i^l, \vec{\mu}_i^l, K_i^l | i = 1, \ldots, M\}, l = 1, \ldots, m. \tag{11}$$

A single model Λ_L of speaker L is obtained by the collection of all the parameters estimated in (11), i. e.,

$$\Lambda_L = \{\pi_i^l, \vec{\mu}_i^l, K_i^l | l = 1, \ldots, m; i = 1, \ldots, M\}. \tag{12}$$

In order to adapt the Colored-MT to the α-GMM classifier, the likelihood $p(\mathbf{x}|\lambda_L)$ is adjusted to follow the α-integration of all $m \times M$ Gaussian densities:

$$p(\mathbf{x}|\Lambda_L) = \mathcal{C}'\left[\sum_{l=1}^{m}\sum_{i=1}^{M} \pi_i^l b_i^l(\mathbf{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}, \tag{13}$$

where $\mathcal{C}'$ is a new normalization constant. As in [18], the Colored-MT is also adopted to obtain the α-GMM for the UBM.

Fig. 3 illustrates a block diagram with the general work-flow of the speaker verification system considered in this work. The highlighted blocks indicate the two techniques, multi-style training and speech enhancement, that are used to improve the performance of speaker verification in noise conditions.

III. Speech Enhancement

The estimation of the noise power spectrum has a major impact on the speech enhancement performance. Since the six acoustic noises considered in this work present highly time-varying spectral characteristics (refer to Section IV-A), the estimation of the noise spectrum should not be restricted to segments where voice is absent. This motivates the use of the MS and IMCRA noise estimators to compose the speech enhancement techniques. The idea is to evaluate the contribution of the MS/SS and IMCRA/OMLSA on reducing the effects of non-stationary noises in speaker verification.

A. MS/SS

The noise estimation based on MS does not need a VAD to distinguish between speech activity and pause phases. The MS is based on the fact that the noisy speech spectrum evaluated over short-time frames often decays to the noise power level, even during speech activity. Thus, an accurate estimation of the noise power at each frequency band can be obtained by searching for the minimum power levels of noisy speech among a few recent past frames.

Let $y(t)$ be a speech utterance corrupted by an additive noise $\eta(t)$. Thus, it can be written $y(t) = x(t) + \eta(t)$, where $x(t)$ represents the clean speech signal. The analysis of $y(t)$ using the short-time Fourier transform (STFT) leads to the following relation

$$Y(\kappa,\tau) = X(\kappa,\tau) + \mathcal{N}(\kappa,\tau), \tag{14}$$

where κ and τ are the frequency bin and the time frame indexes, respectively.

The first step of the MS algorithm is to obtain a recursively smoothed periodogram,

$$S_s(\kappa,\tau) = \delta_s(\kappa,\tau)S_s(\kappa,\tau-1) + (1-\delta_s(\kappa,\tau))|Y(\kappa,\tau)|^2, \tag{15}$$

where $\delta_s(\kappa,\tau)$ is a time- and frequency-dependent smoothing parameter that varies in the range $\delta_s(\kappa,\tau) \in [0,1]$. The noise power is estimated as the minimum values of $S_s(\kappa,\tau)$ obtained from the past $\mathcal{L}$ frames,

$$S_{min}(\kappa,\tau) = \min\left\{S_s(\kappa,\tau');\tau-\mathcal{L} < \tau' \leq \tau\right\}. \tag{16}$$

Since, in general, the minimum value of a random variable is smaller than its average, the noise estimation based on $S_{min}(\kappa,\tau)$ is biased towards smaller values. Hence, a bias compensation factor is needed, and the estimated noise power spectrum is given by

$$|\hat{\mathcal{N}}(\kappa,\tau)|^2 = B_{min}(\kappa,\tau)\cdot S_{min}(\kappa,\tau). \tag{17}$$

In this work, the bias compensation factor $B_{min}(\kappa,\tau)$ in (17) and the optimum value for the smoothing parameter $\delta_s(\kappa,\tau)$ in (15) are obtained following the procedures presented in [27].

After the noise power spectrum estimation, the spectral subtraction is used to reconstruct the enhanced speech. In the SS method [25], the noise power is firstly subtracted from the noisy speech power spectrum,

$$|\bar{X}(\kappa,\tau)|^2 = |Y(\kappa,\tau)|^2 - \xi_{ov}(\kappa,\tau)\cdot|\hat{\mathcal{N}}(\kappa,\tau)|^2, \tag{18}$$

where $\xi_{ov}(\kappa,\tau) \geq 1$ is the oversubtraction factor. The clean speech power spectrum is then estimated as [25]

$$|\hat{X}(\kappa,\tau)|^2 = \begin{cases} |\bar{X}(\kappa,\tau)|^2, & \text{if} |\bar{X}(\kappa,\tau)|^2 > \theta_f\cdot|\hat{\mathcal{N}}(\kappa,\tau)|^2, \\ \theta_f\cdot|\hat{\mathcal{N}}(\kappa,\tau)|^2, & \text{elsewhere}, \end{cases} \tag{19}$$

where $0 < \theta_f \ll 1$ is the spectral floor parameter. Following the procedure in [25], the spectral floor parameter is set to 0.01 and the value of $\xi_{ov}(\kappa,\tau)$ is determined according to the *a posteriori* SNR: the higher the SNR, the lower the oversubtraction factor. The minimum value of $\xi_{ov}(\kappa,\tau)$ is set to 1, corresponding to a SNR of 20 dB.

From (19), the spectrum of the enhanced signal is estimated using the phase $\phi_y(\kappa,\tau)$ of the original speech signal, i. e.,

$$\hat{X}(\kappa,\tau) = |\hat{X}(\kappa,\tau)|e^{j\phi_y(\kappa,\tau)}. \tag{20}$$

Finally, the enhanced speech signal $\hat{x}(t)$ is reconstructed by overlap-adding the inverse Fourier transform of $\hat{X}(\kappa,\tau)$.

B. IMCRA/OMLSA

The second speech enhancement technique applied in this work uses the IMCRA [28] to obtain an estimate of the noise spectrum, and the OMLSA speech estimator [31] to reconstruct the enhanced version of the clean speech. The IMCRA method is composed of two iterations. Firstly, a voice activity detector (VAD) is applied for each frequency bin and time frame. In a second stage, this VAD is used to improve the robustness of the noise tracking during voice activity.

During the first iteration, a noisy spectrum is obtained by frequency and time smoothing:

$$\begin{cases} S_f(\kappa,\tau) = \sum_{i=-w}^{w} W(i)|Y(\kappa-i,\tau)|^2, \\ S(\kappa,\tau) = \delta_s S(\kappa,\tau-1) + (1-\delta_s)S_f(\kappa,\tau), \end{cases} \tag{21}$$

where $\delta_s = 0.9$ is a smoothing parameter and $W(i)$ is a normalized Hanning window function constrained to $\sum_{i=-w}^{w} W(i) = 1$, where $w = 1$. Based on the comparison among the values of $|Y(\kappa,\tau)|$, $S(\kappa,\tau)$ and $S_{min}(\kappa,\tau)$, calculated using (16), a VAD decision criteria is defined for each time and frequency indexes,

$$I(\kappa,\tau) = \begin{cases} 1, & \text{speech absence}, \\ 0, & \text{speech presence}. \end{cases} \tag{22}$$

In the second iteration, a new smoothing spectrum $\tilde{S}(\kappa,\tau)$ is estimated only considering the power spectral components for which the VAD detected primarily noise,

$$\begin{cases} \tilde{S}_f(\kappa,\tau) = \dfrac{\sum_{i=-w}^{w} W(i)I(\kappa-i,\tau)|Y(\kappa-i,\tau)|^2}{\sum_{i=-w}^{w} W(i)I(\kappa-i,\tau)}, \\ \tilde{S}(\kappa,\tau) = \delta_s\tilde{S}(\kappa,\tau-1) + (1-\delta_s)\tilde{S}_f(\kappa,\tau). \end{cases} \tag{23}$$

If the denominator in (23) equals zero, it is replaced by $\tilde{S}(\kappa,\tau) = \tilde{S}(\kappa,\tau-1)$. From the relation between $|Y(\kappa,\tau)|$, $\tilde{S}(\kappa,\tau)$ and its minimum values $\tilde{S}_{min}(\kappa,\tau) = \min\{\tilde{S}(\kappa,\tau');\tau-\mathcal{L} < \tau' \leq \tau\}$, a probability $p(\kappa,\tau) \in [0,1]$ is defined for the speech presence in time and frequency. The noise power spectrum estimation $|\bar{\mathcal{N}}(\kappa,\tau)|^2$ is recursively given by

$$|\bar{\mathcal{N}}(\kappa,\tau+1)|^2 = \tilde{\delta}_\eta(\kappa,\tau)|\bar{\mathcal{N}}(\kappa,\tau)|^2 + [1-\tilde{\delta}_\eta(\kappa,\tau)]|Y(\kappa,\tau)|^2, \tag{24}$$

where $\tilde{\delta}_\eta(\kappa,\tau)$ is a time-varying smoothing parameter that depends on $p(\kappa,\tau)$ and on a constant $\delta_\eta \in [0,1]$

$$\tilde{\delta}_\eta(\kappa,\tau) \triangleq \delta_\eta + (1-\delta_\eta)p(\kappa,\tau). \tag{25}$$

Finally, a bias compensation factor $\xi = 1.47$ is considered to obtain the noise spectrum estimation

$$|\hat{\mathcal{N}}(\kappa,\tau)|^2 = \xi|\bar{\mathcal{N}}(\kappa,\tau)|^2. \tag{26}$$

After the noise spectrum estimation, the OMLSA method reconstructs the enhanced speech signal $\hat{x}(t)$ by minimizing the mean-square error of the log-spectral amplitude, i.e.,

$$E_{min}\left\{\left(\log|X(\kappa,\tau)| - \log|\hat{X}(\kappa,\tau)|\right)^2\right\}, \tag{27}$$

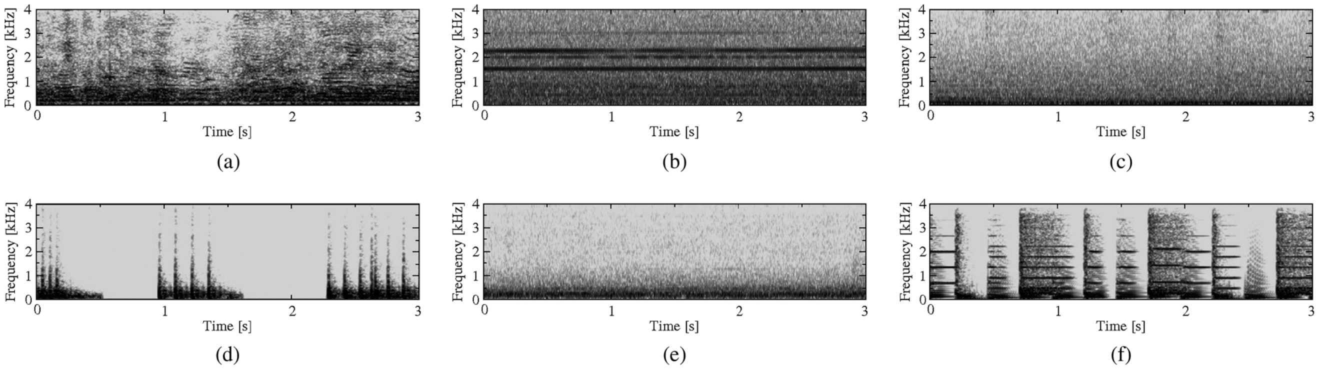


Fig. 4. Spectrograms of the acoustic noises: (a) Babble, (b) Engine, (c) Factory, (d) Machine Gun, (e) Military Vehicle and (f) Ringtone.

where $|\hat{X}(\kappa,\tau)|$ is the spectral amplitude of the optimally reconstructed speech. The power spectrum of the reconstructed signal is calculated by multiplying the spectrum of the noisy speech signal by a gain function,

$$\hat{X}(\kappa,\tau) = G(\kappa,\tau)Y(\kappa,\tau). \tag{28}$$

The gain function $G(\kappa,\tau)$ that leads to the minimum mean-square error in (27) is defined in [31] as

$$G(\kappa,\tau) = \{G_{LSA}(\kappa,\tau)\}^{p(\kappa,\tau)} G_{min}^{1-p(\kappa,\tau)}, \tag{29}$$

where $p(\kappa,\tau)$ is the estimated speech presence probability, $G_{LSA}(\kappa,\tau)$ is a function of the *a priori* SNR at frame τ and frequency bin κ, and the minimum value G_{min} is defined by a subjective criteria. For a detailed description of the OMLSA method, please refer to [31].

IV. Experiments and Results with Non-Stationary Acoustic Noises

The speaker verification experiments presented in this Section are conducted with 158 enrolled speakers collected from the TIMIT database [21]. The speech database is composed of 10 utterances per speaker, with sampling rate of 16 kHz and time average duration of 3 seconds. From each of the enrolled speakers, eight utterances are separated to train the models, and the other two are used for tests, i.e., a total of 316 tests. This leads to 316 genuine or true speaker trials for the false rejection rate evaluation. For the false acceptance evaluation, 49612 impostor trials were used which corresponds to 314 (excluding the speech utterances of the speaker in test) x 158. Regarding the UBM, preliminary experiments are conducted to evaluate the impacts of the UBM composition on the SV results. The EER results are presented and discussed in Section IV-B. They are used to define the number of UBM speakers in the other SV experiments with the TIMIT database.

A. Noise Database

Six acoustic noises (Babble, Engine, Factory, Machine Gun, Military Vehicle and Ringtone) are used to corrupt the speech utterances. The Ringtone noise is available in FindMIDIs.com[1]. The other noises are collected from the NOISEX-92 database [39]. Before being added to the speech utterances, the noises are re-sampled from its original sampling rate (8 kHz for Ringtone and 19.98 kHz for the other five) to 16 kHz. Different values of SNR are considered for the corruption of the speech signals.

Fig. 4 depicts the spectrograms obtained from segments of the six acoustic noises. Note that Babble, Machine Gun and Ringtone noises show high oscillation in their spectrograms. On the other hand, Factory, Engine and Military Vehicle have low spectra variation.

1) Index of Non-Stationarity: In [22], the authors define a process as *stationary relatively to an observation scale* if its local short-time spectra at all different time instants are statistically similar to its global spectrum. The authors proposed the index of non-stationarity as a measure of the time-varying spectra of a random process based on the time-frequency approach. The stationarity test is conducted by comparing the spectral components of the signal to a stationary reference, called *surrogates*.

In this work, the index of non-stationarity (INS) is evaluated following the procedure in [22]. If the noise is stationary, its INS value is expected to be close to unity. On the other hand, the larger the INS the more non-stationary the noise.

The INS values of the 6 acoustic noises are presented in Fig. 5. The time scale T_h/T indicates the relation between the size of the window adopted for the STSA (T_h) and the total length (T) of the noise. The values of γ represent the threshold for the stationarity test, considering a confidence degree of 95%. Thus,

$$\text{INS} \begin{cases} \leq \gamma, & \text{signal is stationary;} \\ > \gamma, & \text{signal is non - stationary.} \end{cases} \tag{30}$$

The comparison between the INS and γ values indicate that the six noises are non-stationary. As expected, the noises whose spectrograms show high oscillation also present the highest values of INS. On the other hand, while the visual inspection of the spectrograms of Factory and Military Vehicle noises indicates low time-varying behavior, the INS of these noises are also above the threshold. This means that they are also non-stationary. Hence, this reinforces the relevance of considering different time scales in the stationarity evaluation. The main issue of using these noises is to evaluate the effect of their non-stationarity on the accuracies of the speaker verification.

B. UBM

In the study presented in [40], it was shown that there is no need for a large number of speakers in the UBM for a conven-

[1] http://www.findmidis.com.

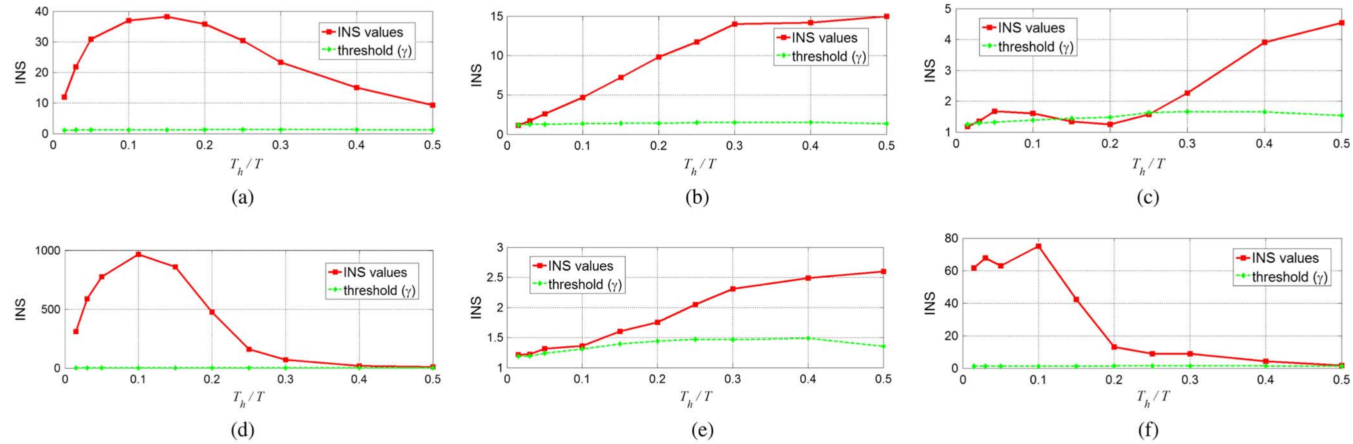


Fig. 5. Indexes of Non-Stationarity of the acoustic noises: (a) Babble, (b) Engine, (c) Factory, (d) Machine Gun, (e) Military Vehicle and (f) Ringtone.

TABLE I
AVERAGE EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH THE MFCC + $\Delta\Delta$ AND THE α-GMM CLASSIFIER FOR THE TIMIT DATABASE

Noise	# speakers in the UBM				
	10	20	40	80	160
Babble	9.48	9.70	9.78	9.70	9.72
Engine	18.76	18.00	18.12	17.87	18.14
Factory	14.27	14.85	14.79	15.02	14.26
Machine Gun	3.40	3.64	3.84	3.60	3.53
Military Vehicle	11.62	12.12	11.72	12.26	11.46
Ringtone	11.54	11.94	11.90	11.85	11.58

tional GMM-UBM SV system. In the present work, a set of preliminary SV experiments based on the α-GMM classifier is used to evaluate the EER with the number of UBM speakers varying from 10 to 160. For this purpose, 12-dimensional MFCC vectors are obtained from frames of 20 ms and 50% of frame overlapping. It is adopted a Mel-scale filterbank with 26 filters and a pre-emphasis factor of 0.97. The MFCC are then appended to their corresponding velocity and acceleration coefficients, leading to MFCC + $\Delta\Delta$ vectors with 36 coefficients.

The speakers (50% male, 50% female) adopted to compose the UBM are randomly chosen from another subset of the TIMIT database. The UBM is obtained from the concatenation of the utterances from all the selected speakers. The UBM and the speakers models are composed with 32 Gaussian densities and four values of α: -1, -4, -6 and -8. The average EER results, considering tests with the four values of SNR (5, 10, 15 and 20 dB), are shown in Table I. These results correspond to the values of α that lead to the lowest average EER for each UBM composition: $\alpha = -4$ for 20 and 80 speakers, $\alpha = -6$ for 10 and 160, and $\alpha = -8$ for 40. Note that there is no significant improvement in the average EER when larger numbers of speakers are adopted for the UBM with the TIMIT database. Due to such results and to large amount of SV experiments conducted in this work, the number of UBM speakers in all the remaining SV experiments conducted with the TIMIT database is set to 10 (5 male and 5 female).

C. Experiments with MFCC + pH Fusion

In the first set of experiments, the speaker verification task is evaluated with the α-GMM considering the MFCC and also the fusion of the MFCC and pH speech features. The experiments are repeated considering the velocity and acceleration coefficients of the MFCC (MFCC + $\Delta\Delta$). None of the multi-style training and speech enhancement techniques are used in these experiments.

The pH are estimated from two consecutive speech frames using Daubechies wavelets filters [35] with 12 detail coefficients and scale range from 3 to 9. As in [18], a total of $J = 8$ decomposition scales are considered to obtain the H_j values. Thus, including the H_0 component obtained from the speech signal, these 9-dimensional vectors are DCT-transformed to compose the pH feature matrices. Thus, for experiments with the MFCC + pH fusion, feature vectors are formed by 21 components. The MFCC + $\Delta\Delta$ and MFCC + $\Delta\Delta$ + pH vectors are composed of 36 and 45 coefficients, respectively.

Table II presents the EER results for the experiments conducted with test utterances corrupted by the six acoustic noises and also for the clean speech. Note that, for the three compositions of the feature vectors (MFCC + $\Delta\Delta$, MFCC + pH and MFCC + $\Delta\Delta$ + pH), the best accuracy, i.e., the lowest average EER, is obtained with $\alpha = -6$. For this value of α, the MFCC + $\Delta\Delta$ + pH fusion achieves the best average performance for 5 of the noise sources. This leads to an absolute improvement of 1.30% in the average EER results, from 11.51% with MFCC + $\Delta\Delta$, to 10.21%. The use of pH feature vectors reduces the EER results in 4.44% for Factory noise and SNR of 5 dB, from 26.27% to 21.83%. In comparison to the conventional MFCC- and GMM-based system, also included in Table II, an absolute average EER reduction of 2.58% is achieved with the MFCC + $\Delta\Delta$ + pH and α-GMM. It can also be seen from Table II that, in comparison to MFCC + $\Delta\Delta$, the MFCC + pH fusion shows the best average accuracies for 4 of the 6 noise sources. The average EER is reduced from 11.51% to 10.67% for $\alpha = -6$. This means that, in comparison to the velocity and acceleration coefficients, the use of pH leads to an overall improvement in the SV task even with a lower number of components in the feature vectors.

TABLE II
EER (%) Obtained in the Speaker Verification Tests with the α-GMM Classifier for Different Values of α

Noise	SNR	MFCC	MFCC + ΔΔ				MFCC + pH				MFCC + ΔΔ + pH			
		$\alpha=-1$	$\alpha=-1$	$\alpha=-4$	$\alpha=-6$	$\alpha=-8$	$\alpha=-1$	$\alpha=-4$	$\alpha=-6$	$\alpha=-8$	$\alpha=-1$	$\alpha=-4$	$\alpha=-6$	$\alpha=-8$
Clean		1.48	1.27	1.27	1.16	1.18	1.28	1.59	1.27	1.59	1.27	1.27	1.41	1.17
Babble	20 dB	2.85	2.22	2.96	2.32	2.22	2.31	3.01	2.47	2.80	2.70	2.76	2.53	2.75
	15 dB	5.06	4.15	5.12	4.11	5.06	4.43	5.06	3.80	4.43	5.24	5.12	5.19	6.07
	10 dB	11.20	11.39	10.44	9.98	11.08	10.44	11.08	10.38	10.88	11.73	11.28	10.76	11.82
	5 dB	25.00	20.57	21.70	21.52	22.78	21.84	22.47	20.57	21.52	21.52	20.89	20.76	21.20
	Average	11.03	9.58	10.06	9.48	10.28	9.75	10.40	9.30	9.91	10.30	10.01	9.81	10.46
Engine	20 dB	4.84	5.54	5.42	5.89	5.70	5.09	6.01	5.46	5.78	6.86	6.75	5.92	6.12
	15 dB	12.14	12.88	12.22	12.82	12.74	12.97	12.90	12.25	12.98	12.57	13.87	12.34	13.29
	10 dB	23.70	22.86	22.15	23.10	23.19	23.73	25.00	23.73	23.03	21.84	22.47	21.84	23.00
	5 dB	37.16	33.62	32.66	33.23	33.39	35.79	37.97	35.76	35.01	32.91	33.50	32.86	33.23
	Average	19.46	18.72	18.11	18.76	18.75	19.40	20.47	19.30	19.20	18.55	19.15	18.24	18.91
Factory	20 dB	5.04	4.02	4.17	4.22	4.75	4.11	4.11	3.93	4.29	4.26	3.61	3.87	4.50
	15 dB	10.13	9.18	9.85	9.46	9.81	7.18	7.33	7.73	7.39	7.97	7.28	7.88	7.59
	10 dB	19.94	17.54	17.92	17.13	19.30	13.61	14.56	14.24	13.67	13.92	13.64	13.92	13.58
	5 dB	30.98	26.58	26.69	26.27	26.27	23.73	24.37	23.10	23.63	21.86	22.94	21.83	21.84
	Average	16.52	14.33	14.66	14.27	15.03	12.16	12.59	12.25	12.25	12.00	11.87	11.87	11.88
Machine Gun	20 dB	2.09	1.58	1.90	1.58	1.58	2.22	2.33	1.90	2.53	2.27	2.14	1.66	1.84
	15 dB	2.92	2.67	2.53	2.22	2.49	2.76	2.85	2.53	3.08	2.74	2.80	2.07	2.74
	10 dB	5.06	3.56	4.11	3.48	3.80	3.42	3.48	3.94	3.82	3.16	3.80	3.78	3.80
	5 dB	7.91	5.86	5.93	6.33	5.92	5.58	5.70	6.01	4.89	5.38	5.06	5.06	5.38
	Average	4.50	3.42	3.62	3.40	3.45	3.49	3.59	3.60	3.58	3.39	3.45	3.14	3.44
Military Vehicle	20 dB	4.43	4.20	4.43	3.46	3.48	2.89	3.43	3.16	3.22	3.16	3.16	3.22	3.80
	15 dB	8.35	7.36	7.28	7.91	7.91	5.88	6.01	6.01	6.65	6.30	5.95	6.33	6.33
	10 dB	14.92	14.01	14.24	13.92	14.71	11.71	12.32	11.73	11.42	10.79	10.91	10.76	11.08
	5 dB	23.92	21.52	21.20	21.19	22.15	18.99	20.25	19.62	19.45	18.67	18.64	17.12	19.94
	Average	12.91	11.77	11.79	11.62	12.06	9.87	10.50	10.13	10.18	9.73	9.67	9.36	10.28
Ringtone	20 dB	3.80	4.11	4.19	4.16	4.11	3.45	3.78	3.16	3.34	3.56	3.63	3.80	4.00
	15 dB	9.03	7.98	8.23	7.86	7.91	6.13	6.35	6.01	6.01	6.33	6.01	6.61	6.18
	10 dB	14.87	14.24	14.32	13.54	14.97	11.39	10.85	11.43	10.87	10.76	10.06	9.81	10.13
	5 dB	21.64	20.89	20.87	20.62	20.98	17.52	16.77	17.09	17.09	16.14	15.27	15.19	15.51
	Average	12.33	11.81	11.90	11.54	11.99	9.62	9.44	9.42	9.33	9.20	8.74	8.85	8.95
Average		12.79	11.61	11.69	11.51	11.93	10.72	11.17	10.67	10.74	10.53	10.48	10.21	10.65

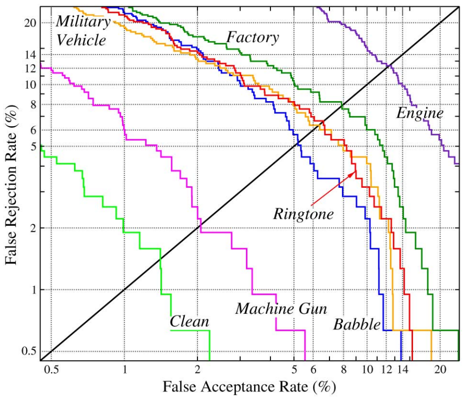


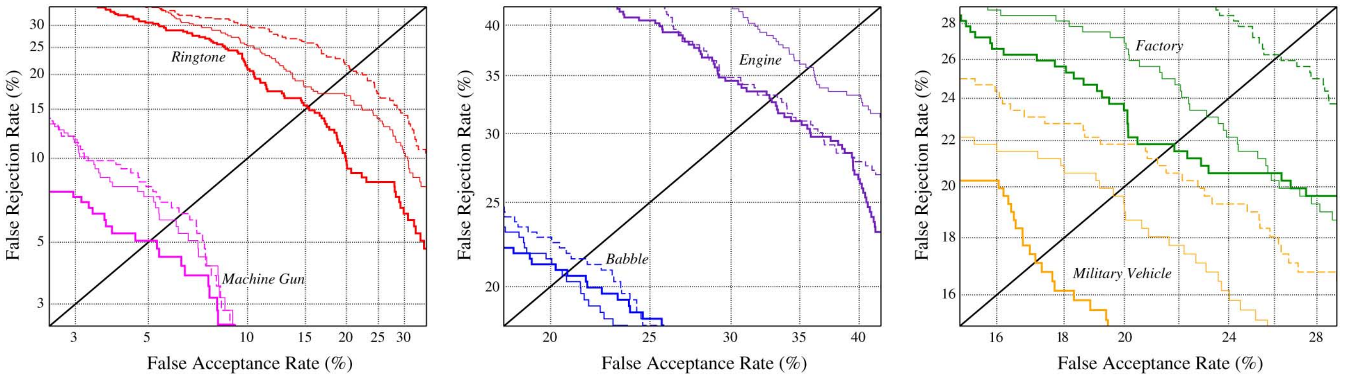
Fig. 6. The DET curves obtained with MFCC + ΔΔ + pH with the α-GMM classifier ($\alpha = -6$) for test speech signals corrupted by the acoustic noises with SNR of 15 dB, and also for clean speech.

The DET curves from the experiments conducted with the MFCC + ΔΔ + pH fusion, $\alpha = -6$ and noise corruption with SNR of 15 dB, and also for the clean speech, are illustrated in Fig. 6. The EER values are represented by the operating points where the DET curves cross the black line.

Fig. 7 depicts the DET curves considering the MFCC + ΔΔ, the MFCC + pH and the MFCC + ΔΔ + pH vectors, considering $\alpha = -6$. The curves are presented according to their indexes of non-stationarity. While the left curves are related to noises with the highest INS, those on the right are obtained for noises with the lowest INS. In these experiments, the noises corruption adopts SNR of 5 dB. These results reinforce the improvement in the verification accuracies due to the use of the MFCC and pH feature fusion in severe noise conditions. Due to the average improvement obtained with the pH, the MFCC + ΔΔ + pH features fusion is adopted in all the following experiments.

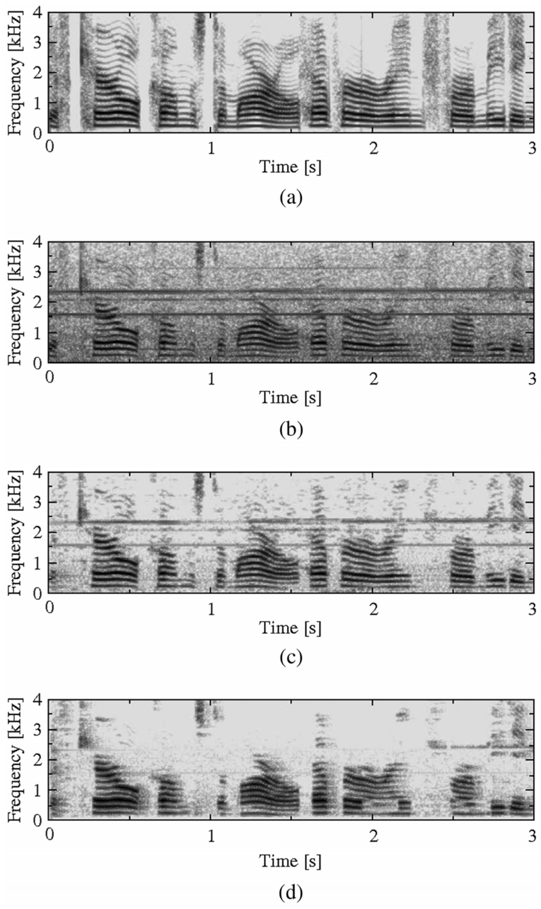
D. Experiments with Speech Enhancement

In the second set of experiments, the MS/SS and IMCRA/OMLSA speech enhancement techniques are applied as preprocessing steps to the SV task (refer to Fig. 3). For both techniques, the noisy speech is split into 50%-overlapping frames with length of 512 samples. The speaker models are obtained without any enhancement. Fig. 8 illustrates the suppression of



(a) (b) (c)

Fig. 7. DET curves obtained from experiments with MFCC + $\Delta\Delta$ (dashed lines), MFCC + pH (thin continuous lines) and MFCC + $\Delta\Delta$ + pH (thick continuous lines): (a) Machine Gun and Ringtone, (b) Engine and Babble and (c) Factory and Military Vehicle.



(a)

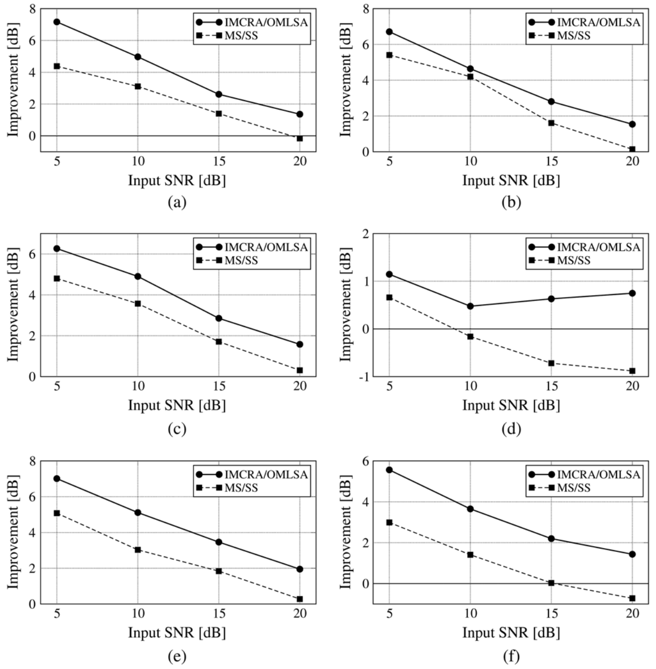
(b)

(c)

(d)

Fig. 8. Spectrograms from a male speaker: (a) clean speech; (b) speech corrupted by Engine noise with SNR of 10 dB and enhanced speech with (c) MS/SS and (d) IMCRA/OMLSA.

noise from a speech segment corrupted by Engine noise with SNR of 10 dB. The comparison among the spectrograms of clean speech, noisy signal and the enhanced speech shows that both MS/SS and IMCRA/OMLSA techniques are able to suppress most of the high energy frequencies of the Engine noise. However, some of the spectral components present in the clean speech are also removed by the enhancement techniques.



(a) (b) (c) (d) (e) (f)

Fig. 9. The average SegSNR improvements (dB) obtained with the speech enhancement techniques for different noise sources: (a) Babble, (b) Engine, (c) Factory, (d) Machine Gun, (e) Military Vehicle and (f) Ringtone.

The segmental SNR (SegSNR) is adopted to objectively measure the performance of the speech enhancement. The SegSNR is defined as

$$\mathrm{SegSNR} = \frac{10}{|\mathcal{S}|}\sum_{\tau\in\mathcal{S}}\log\frac{\sum_{\kappa}|X(\kappa,\tau)|^2}{\sum_{\kappa}|\mathcal{N}(\kappa,\tau)|^2}, \tag{31}$$

where X and $\mathcal{N}$ are the STFT components defined in (14), $\mathcal{S}$ is the set of frames that contain voice and $|\mathcal{S}|$ is its cardinality.

Fig. 9 compares the average SegSNR improvement results obtained with the MS/SS and IMCRA/OMLSA techniques for the six acoustic noises and for different input SNR values. Note that IMCRA/OMLSA achieves positive gain for all noise conditions. For the MS/SS, the average SegSNR is not improved for the 3 noises with highest values of INS, particularly for the less severe noise levels.

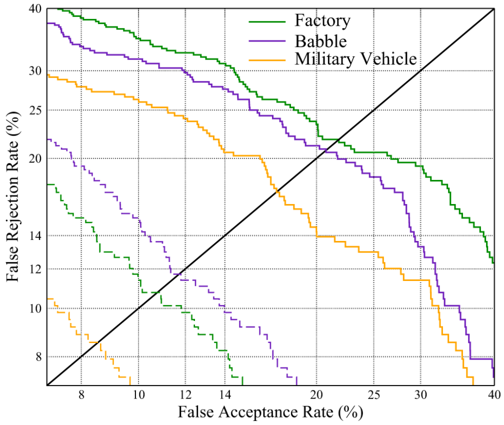


Fig. 10. The DET curves from experiments with SNR of 5 dB using α-GMM. Dashed curves are obtained with MS/SS and $\alpha = -4$. Continuous curves are from SV considering $\alpha = -6$ without any speech enhancement.

The lowest SegSNR improvement, considering both techniques, is obtained for the Machine Gun noise. The worse performance is due to the fact that, besides the highest INS values, the Machine Gun also presents the characteristics of an impulsive noise (refer to its spectrogram in Fig. 4). As discussed in [41], although the MS and IMCRA techniques estimate the noise power every time frame, they can present inaccurate results on tracking sudden and abrupt changes in the noise spectrum.

Table III shows the EER results from SV experiments with test utterances enhanced by both techniques. The results correspond to the values of α that led to the best EER results, which correspond to $\alpha = -4$ for MS/SS and $\alpha = -6$ for IMCRA/OMLSA. Note that, except for the Engine noise, the EER values obtained with MS/SS are much lower than those obtained with IMCRA/OMLSA.

When compared to the results obtained without the use of a speech enhancement technique (MFCC + $\Delta\Delta$ + pH with $\alpha = -6$ in Tab. II), the adoption of the MS/SS leads to a lower average EER for 5 of the noise sources. The absolute EER reduction is 11.82% for the Engine and 11.07% for the Factory noise, both for SNR of 5 dB. Fig. 10 illustrates the contribution of the MS/SS speech enhancement on reducing the FA and FR errors for 3 different noises: Babble, Factory and Military Vehicle noises. The only noise for which the EER is not reduced is Machine Gun. As previously discussed, this is due to the difficulty of the MS estimator on tracking the power spectra of impulsive noises.

E. Experiments with Multi-Style Training

In the third set of experiments, the colored-noise-based MT [17] is applied to improve the robustness of the speaker verification system. Following the procedure defined in [17], three artificial noises are generated for the multi-style training, with colored spectra defined by the PSD decaying rate: $\beta = 0$ (white), $\beta = 1$ (pink) and $\beta = 2$ (brown). These noises are used to corrupt all the speech segments available for training with SNR of 15 dB, including the UBM. Thus, a total of $3 \times 32 = 96$ Gaussian densities are stored for each speaker.

Two other multi-style training techniques are used as references for the Colored-MT. A white-noise-based MT (White-MT) [8] is obtained by corrupting multiple copies of the training utterances with SNR values between 10 and 20 dB, with step of 2 dB[2]. Following the procedure in [8], the clean and corrupted training utterances are then concatenated and used to train the UBM and speaker models with 128 Gaussian components. The same procedure is adopted to obtain the models in the narrow-band-noise-based MT (Narrow-MT). The narrow-band noise is obtained by passing the white noise through a low-pass filter with a lower 3-dB cutoff frequency of 800 Hz [8].

TABLE III
EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH MFCC + $\Delta\Delta$ + PH FEATURES WITH α-GMM CLASSIFIER AND SPEECH ENHANCEMENT

Noise	SNR	MS/SS ($\alpha = -4$)	IMCRA/OMLSA ($\alpha = -6$)
Babble	20 dB	2.74	8.37
	15 dB	3.80	9.49
	10 dB	6.96	10.76
	5 dB	11.71	15.19
	Average	6.30	10.95
Engine	20 dB	3.64	8.31
	15 dB	7.40	9.10
	10 dB	13.29	10.44
	5 dB	21.04	15.43
	Average	11.34	10.82
Factory	20 dB	2.90	7.91
	15 dB	3.80	7.59
	10 dB	6.33	8.69
	5 dB	10.76	10.01
	Average	5.95	8.55
Machine Gun	20 dB	3.16	9.45
	15 dB	3.16	9.81
	10 dB	3.81	10.76
	5 dB	6.10	12.86
	Average	4.06	10.72
Military Vehicle	20 dB	2.93	10.13
	15 dB	3.80	11.30
	10 dB	6.33	12.04
	5 dB	8.54	12.03
	Average	5.40	11.37
Ringtone	20 dB	3.64	9.46
	15 dB	5.62	10.21
	10 dB	9.18	12.97
	5 dB	13.78	18.04
	Average	8.05	12.67
Average		6.85	10.85

SV experiments are conducted with the MT techniques and α-GMM considering the four values of α: -1, -4, -6 and -8.

[2] The SNR range of 10-20 dB is adopted in this work since it led to better results than the SNR range 4-20 dB adopted in [8].

Table IV presents the results corresponding to the value of α that led to lowest average EER. Note that the Colored-MT with $\alpha = -6$ achieves the best results for 5 of the 6 noises. For Factory noise with SNR of 5 dB, the Colored-MT reduces the EER from 24.05% with White-MT and 18.34% with Narrow-MT to 6.34%. In average, the Colored-MT achieves absolute overall reduction of 3.14% and 5.04% in comparison to the Narrow-MT and the White-MT, respectively[3].

The results in Table IV also show that the Colored-MT achieves lower average EER values than those obtained with speech enhancement (refer to Table III) for 5 noise sources. The average EER is reduced from 6.85% (MS/SS in Table III) to 6.31%. The Engine noise is the only one for which the adoption of MS/SS speech enhancement outperforms the Colored-MT. This fact can be explained by the high energy frequencies between 1.5 kHz and 2.5 kHz in the spectrum of the Engine noise (see Fig. 4(b)). It means that the energy of Engine noise is not concentrated at low-frequency components, as is the case of colored noises. On the other hand, as shown in Fig. 8, the noise components related to these frequencies are removed by the speech enhancement techniques. Thus, the results obtained with MS/SS and IMCRA/OMLSA are better than those obtained with Colored-MT for this specific noise source.

It may also be noticed that, when compared to the results obtained without speech enhancement (MFCC + $\Delta\Delta$ + pH and $\alpha = -6$ in Table II), the use of Colored-MT improves the average performance of SV for all noise sources. The absolute EER reduction achieves 15.49% for Factory noise with SNR of 5 dB, from 21.83% to 6.34%. For the highly non-stationary Ringtone noise the average EER is reduced from 8.85% to 5.54%. Even for the Machine Gun noise, which presents the highest values of INS, the performance is improved with the Colored-MT.

F. Experiments with Speech Enhancement and Multi-Style Training

In the fourth set of experiments, both the Colored-MT and the MS/SS speech enhancement are applied to improve the robustness of the SV. For the multi-style training, the same artificial (white, pink and brown) noises are used to corrupt the training utterances and obtain the speaker models. Before the extraction of the MFCC + $\Delta\Delta$ + pH feature matrices, all the training and test utterances are enhanced using the MS/SS technique. The MS/SS is chosen due its significant improvement in the EER results, when compared to the IMCRA/OMLSA technique.

The EER results obtained with both the MS/SS and the Colored-MT are presented in the last column of Table V. These results correspond to $\alpha = -8$, which leads to an average EER slightly lower than those obtained with the other values of α. For comparison, the lowest EER results obtained without any techniques (Table II), with the MS/SS only (Table III) and the Colored-MT only (Table IV) are also shown in Table V. It can be seen that the adoption of both techniques improves the performance of the speaker verification for two acoustic noises: Engine and Military Vehicle. The overall EER result is reduced from 6.31% (with Colored-MT only) to 5.84%.

TABLE IV
EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH MFCC + $\Delta\Delta$ + PH FEATURES AND MULTI-STYLE TRAINING WITH α-GMM CLASSIFIER

Noise	SNR	White-MT ($\alpha = -4$)	Narrow-MT ($\alpha = -4$)	Colored-MT ($\alpha = -6$)
Babble	20 dB	4.46	4.11	2.92
	15 dB	6.88	5.38	3.73
	10 dB	12.24	8.58	5.57
	5 dB	23.42	16.29	10.11
	Average	11.75	8.59	5.58
Engine	20 dB	6.01	5.14	7.21
	15 dB	11.48	8.86	11.71
	10 dB	19.94	15.86	17.72
	5 dB	30.76	27.00	28.45
	Average	17.05	14.21	16.27
Factory	20 dB	5.35	4.43	1.84
	15 dB	8.86	6.76	2.22
	10 dB	15.10	11.16	3.03
	5 dB	24.05	18.34	6.34
	Average	13.34	10.17	3.36
Machine Gun	20 dB	3.16	3.16	2.53
	15 dB	3.80	3.96	2.62
	10 dB	4.74	5.06	3.16
	5 dB	6.74	6.65	3.48
	Average	4.61	4.71	2.95
Military Vehicle	20 dB	4.43	4.01	2.16
	15 dB	7.66	6.39	2.97
	10 dB	13.29	10.90	4.35
	5 dB	19.94	18.67	7.20
	Average	11.33	9.99	4.17
Ringtone	20 dB	4.66	4.19	2.85
	15 dB	7.28	6.33	3.48
	10 dB	11.71	10.13	5.38
	5 dB	16.46	15.51	10.44
	Average	10.03	9.04	5.54
Average		11.35	9.45	6.31

It may be observed that the contribution of the MS/SS for Engine noise is noticeable: the average EER is reduced from 16.27% to 9.97%. This noteworthy improvement can be explained by, as discussed in Section IV-E, the presence of the energy peaks in the Engine noise spectrum that are suppressed by the MS/SS speech enhancement. On the other hand, the Colored-MT presents only a slight contribution to SV robustness for this specific noise, since its energy is not concentrated in the low-frequency part of the spectrum (refer to the spectrograms in Fig. 4).

Regarding the average EER results obtained for each noise, note that the MT without MS/SS achieves the best performance for the 3 highly non-stationary noises (Babble, Machine Gun

[3] It is important to mention that, different from the experiments presented in [8], in this work the evaluation of the multi-style training techniques does not consider the use of subband features.

TABLE V
EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH MFCC + $\Delta\Delta$ + PH FEATURES, MS/SS AND COLORED-MT WITH α-GMM CLASSIFIER

Noise	SNR	without any technique	MS/SS	Colored-MT	MS/SS and Colored-MT
Babble	20 dB	2.53	2.74	2.92	3.94
	15 dB	5.19	3.80	3.73	4.12
	10 dB	10.76	6.96	5.57	6.10
	5 dB	20.76	11.71	10.11	11.39
	Average	9.81	6.30	5.58	6.39
Engine	20 dB	5.92	3.64	7.21	4.75
	15 dB	12.34	7.40	11.71	6.65
	10 dB	21.84	13.29	17.72	10.38
	5 dB	32.86	21.04	28.45	18.11
	Average	18.24	11.34	16.27	9.97
Factory	20 dB	3.87	2.90	1.84	3.38
	15 dB	7.88	3.80	2.22	3.42
	10 dB	13.92	6.33	3.03	3.93
	5 dB	21.83	10.76	6.34	6.81
	Average	11.87	5.95	3.36	4.39
Machine Gun	20 dB	1.66	3.16	2.53	3.73
	15 dB	2.07	3.16	2.62	3.59
	10 dB	3.78	3.81	3.16	4.40
	5 dB	5.06	6.10	3.48	4.43
	Average	3.14	4.06	2.95	4.04
Military Vehicle	20 dB	3.22	2.93	2.16	2.85
	15 dB	6.33	3.80	2.97	3.16
	10 dB	10.76	6.33	4.35	4.11
	5 dB	17.12	8.54	7.20	5.40
	Average	9.36	5.40	4.17	3.88
Ringtone	20 dB	3.80	3.64	2.85	4.11
	15 dB	6.61	5.62	3.48	4.79
	10 dB	9.81	9.18	5.38	6.48
	5 dB	15.19	13.78	10.44	10.13
	Average	8.85	8.05	5.54	6.38
Average		10.21	6.85	6.31	5.84

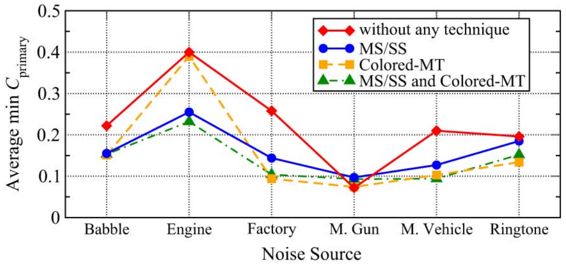


Fig. 11. The average min C_{primary} results obtained in SV experiments with the six noises.

and Ringtone). It may also be noted that, for all the noise sources, the lowest EER results are obtained with the Colored-MT (with and without the speech enhancement).

As a complement to the EER, the minimum value of the primary cost defined in [32] is also used to measure the SV performance. The average min C_{primary} results for each noise in the four sets of experiments are depicted in Fig. 11. In agreement with the EER results (Table V), note that the adoption of MS/SS and Colored-MT leads to the best results for the Engine and the Military Vehicle noises. For the Factory and Ringtone noises, the lowest min C_{primary} results are achieved with the multi-style training only, while the speech enhancement and the multi-style training achieve similar average results for the Babble noise. Finally, Machine Gun is the only noise source for which the SV without any technique achieved the best SV performance, with average result similar to that obtained with MT only.

V. EXPERIMENTS AND RESULTS IN REALISTIC NOISY ENVIRONMENTS

The contribution of the speech enhancement and the multi-style training for SV is also evaluated in realistic noisy conditions. For this purpose, SV experiments are also conducted with the MIT Mobile Device Speaker Verification Corpus [23]. The MIT database is composed of 48 enrolled speakers and 40 impostors. The speech signals were collected with a handheld-device using an internal microphone and an external headset in three different environments: an office with low background noise level, a mildly noisy lobby and a street intersection with high background noise level. By using this database, the Lombard effect is also taken into account in the SV experiments. In the text-independent SV experiments here described, it is adopted the subset of the database that corresponds to all the lists of names. It means that, for each enrolled speaker and each test condition, 5 utterances corresponding to spoken names are used for training and other 5 are available for the tests. The utterances from the impostors are used to obtain the UBM. This leads to $48 \times 5 = 240$ genuine trials and $235 \times 48 = 11280$ impostor trials. Only the utterances recorded in the office environment were adopted for training the speakers models and the UBM.

Table VI presents the EER results obtained with α-GMM and three sets of speech features: MFCC, MFCC + $\Delta\Delta$ and MFCC + $\Delta\Delta$ + pH. Both the MFCC and the pH are extracted in the same manner as in Section IV. The value $\alpha = -8$ achieves the lowest average EER results when compared to the other values. For comparison, the EER results achieved with the conventional GMM ($\alpha = -1$) and the MFCC are also shown in Table VI. When only the external headset is considered, the MFCC + $\Delta\Delta$ + pH features fusion leads to the lowest EER values for the three different environments. Regarding the adoption of the internal microphone, the MFCC + $\Delta\Delta$ and the MFCC + $\Delta\Delta$ + pH features sets achieve similar average performance: 26.2% for the former and 26.3% for the latter.

The MT techniques are also examined for the SV experiments with the MIT database. The EER results obtained with the White-MT, the Narrow-MT and the Colored-MT with α-GMM and the MFCC + $\Delta\Delta$ + pH features are presented in Table VII. Once again, the lowest EER results are obtained with the value $\alpha = -8$. Note that, when compared to the results without the multi-style training (Table VI), the Colored-MT improves the

TABLE VI
EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH THE α-GMM CLASSIFIER FOR THE MIT DATABASE

Feature	α	External Headset			Internal Microphone		
		Street	Lobby	Office	Street	Lobby	Office
MFCC	-1	40.8	18.2	12.5	36.3	25.7	20.1
	-8	38.3	18.3	11.8	35.0	27.1	19.2
MFCC + $\Delta\Delta$	-8	37.9	16.7	10.4	32.6	26.7	19.2
MFCC + $\Delta\Delta$ + pH	-8	37.0	15.7	9.6	32.9	26.3	19.6

TABLE VII
EER (%) OF SPEAKER VERIFICATION EXPERIMENTS WITH MFCC + $\Delta\Delta$ + pH FEATURES, MULTI-STYLE TRAINING WITH THE α-GMM CLASSIFIER AND SPEECH ENHANCEMENT TECHNIQUES FOR THE MIT DATABASE

Applied Techniques	α	External Headset			Internal Microphone		
		Street	Lobby	Office	Street	Lobby	Office
White-MT	-8	36.3	18.5	12.1	34.4	27.5	20.0
Narrow-MT	-8	38.3	17.9	12.1	34.2	25.0	21.2
Colored-MT	-8	31.4	15.0	8.8	30.4	24.2	18.3
MS/SS	-6	24.8	15.4	10.1	31.5	23.1	16.7
MS/SS and Colored-MT	-6	34.6	12.5	10.4	29.6	19.7	15.5

TABLE VIII
THE AVERAGE MIN C_{PRIMARY} RESULTS OBTAINED IN SV EXPERIMENTS WITH THE MIT DATABASE

Applied Techniques	α	External Headset			Internal Microphone		
		Street	Lobby	Office	Street	Lobby	Office
no technique	-8	0.738	0.347	0.232	0.714	0.568	0.440
Colored-MT	-8	0.689	0.316	0.213	0.659	0.540	0.404
MS/SS	-6	0.550	0.351	0.236	0.664	0.498	0.377
MS/SS and Colored-MT	-6	0.722	0.287	0.239	0.659	0.436	0.355

SV performance for all the three environments and the two microphones. For instance, the EER for the external headset and the office environment is reduced from 9.6% to 8.8%. Moreover, it outperforms the White-MT and the Narrow-MT for all the six conditions.

The results with the MS/SS speech enhancement technique are also presented in Table VII. The lowest average EER results (with and without MT) are obtained with $\alpha = -6$. For the experiments considering the external headset and the Street environment, the MS/SS leads to an absolute EER reduction of 12.2%, from 37.0%, without any technique, to 24.8%. In this specific condition, the SV with the MS/SS outperforms even the results achieved with the MS/SS and the Colored-MT. It occurs due to the background noise in the Street environment of MIT database seems to be stationary. Thus, the MS/SS technique is able to suppress most part of the Street noise from the speech signals and, consequently, reduce the mismatch between the training and test phases. For the Lobby environment, the best performance is achieved with the MS/SS and the Colored-MT. For this scenario, the EER is reduced from 15.7% to 12.5%, which means an absolute reduction of 3.2%. Regarding the adoption of the internal microphone, the use of the MS/SS and the multi-style training achieves the best results for all the three environments.

The average min C_{primary} results obtained with MS/SS and Colored-MT are shown in Table VIII. In agreement with the EER results (Table VII), the combination of speech enhancement with the multi-style training leads to the best performance for the three environments considering the internal microphone. Regarding the experiments with the external headset, the lowest min C_{primary} results in the Street, Office and Lobby environments are achieved with MS/SS, Colored-MT and MS/SS + Colored-MT, respectively.

VI. CONCLUSION

This paper examined the fusion use of the MFCC and pH features for noise robust speaker verification. The α-GMM classifier was adopted for the speakers and UBM modeling. The experiments were firstly conducted with a subset of the TIMIT database corrupted with six non-stationary acoustic noises and different values of SNR. The index of non-stationarity of these noises were also evaluated in this work. Then, the SV was also evaluated in realistic noisy conditions using the MIT database. The SV results showed that the use of pH features reduced the average EER results obtained with MFCC and their corresponding velocity and acceleration coefficients. The lowest average EER were obtained with the α-GMM classifier with $\alpha = -6$ and $\alpha = -8$ for the TIMIT and MIT databases, respectively. The speaker verification experiments were repeated with Speech enhancement and multi-style training techniques were also evaluated to improve the speaker verification results. Experiments with MFCC + $\Delta\Delta$ + pH features, $\alpha = -6$, Colored-MT and MS/SS achieved the best overall EER results for both databases. The minimum value of the C_{primary} measure was also adopted to reinforce the efficiency of the speech enhancement and MT techniques. Finally, the speech enhancement and the multi-style training showed to be good solutions to improve the SV performance in different noisy conditions.

REFERENCES

[1] J. Naik, "Speaker verification: A tutorial," *IEEE Commun. Mag.*, pp. 42–48, Jan. 1990.

[2] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag* , vol. 26, no. 2, pp. 95–103, Mar. 2009.

[3] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–82, Jan. 1995.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[5] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 4, pp. 430–451, 2004.

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[7] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[8] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[9] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '06)*, May 2006, vol. 1, pp. 97–100.

[10] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '05)*, Mar. 2005, vol. 1, pp. 629–632.

[11] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[12] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[14] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV '07)*, Oct. 2007, pp. 1–8.

[15] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1987, vol. 12, pp. 705–708.

[16] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, 2000, pp. 29–32.

[17] L. Zão and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 675–678, Nov. 2011.

[18] L. Zão and R. Coelho, "Noise robust speaker verification based on the MFCC and pH features fusion and multicondition training," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process. (BIOSIGNALS '12)*, Feb. 2012, pp. 137–143.

[19] D. Wu, J. Li, and H. Wu, "α-Gaussian mixture modelling for speaker recognition," *Pattern Recogn. Lett.*, vol. 30, no. 6, pp. 589–594, 2009.

[20] R. Sant'Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 931–940, May 2006.

[21] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recogn.*, 1986, pp. 93–99.

[22] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3459–3470, Jul. 2010.

[23] R. Woo, A. Park, and T. Hazen, "The MIT mobile device speaker verification corpus: Data collection and preliminary experiments," in *Proc. Odyssey Speaker Lang. Recogn. Workshop*, Jun. 2006, pp. 1–6.

[24] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf, Acoust., Speech, Signal Process. (ICASSP '79)*, Apr. 1979, vol. 4, pp. 208–211.

[26] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. 24th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '99)*, Mar. 1999, pp. 789–792.

[27] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[28] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[29] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. IEEE Int. Conf, Acoust., Speech, Signal Process. (ICASSP)*, May 1998, pp. 121–124.

[30] J. Ortega-Garcï¿½a and J. Gonzï¿½lez-Rodrï¿½guez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 929–932.

[31] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[32] "The NIST year 2012speaker recognition evaluation plan," [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm 2012

[33] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[34] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 878–897, Apr. 1999.

[35] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Ind. and Appl. Math., 1992.

[36] M. Vetterli and J. Kovacevic, *Wavelets and subband coding*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.

[37] D. Wu, "Parameter estimation for α-GMM based on maximum likelihood criterion," *Neural Comput.*, vol. 21, no. 6, pp. 1776–1795, 2009.

[38] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-gaussian distribution," *IET Signal Process.*, vol. 6, no. 7, pp. 684–688, Sep. 2012.

[39] A. Varga and H. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[40] T. Hasan and J. Hansen, "A study on universal background model training in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1890–1899, Feb. 2011.

[41] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Commun.*, vol. 48, pp. 96–109, Jan. 2006.

André Bastos Venturini obtained the M.Sc. degree in electrical engineering from the Military Institute of Engineering (IME) of Rio de Janeiro in 2011. He received the B.Sc. degrees in electrical engineering in 2004 from the Ecole Centrale de Lyon, in France, and also from the Catholic University of Rio de Janeiro (PUC-Rio). His current research activities are mainly related to speaker recognition.

Leonardo Augusto Zão obtained the Ph.D. degree from the Military Institute of Engineering (IME) of Rio de Janeiro in 2013. From the same Institute, he received the M.Sc. and B.Sc. degrees in electrical engineering in 2010 and 2005, respectively. Since 2014, he works at the Laboratory of Acoustic Signal Processing (LASP) as Research Assistant. His current research mainly focuses on speaker recognition, speech enhancement, speech emotion classification, and acoustic signal processing.

Rosângela Fernandes Coelho received the Ph.D. degree from the Ecole Nationale Supérieure des Télécommunications (ENST-Paris) in 1995 and the M.Sc. degree from the Pontifícia Universidade Católica of Rio de Janeiro (PUC-Rio) in 1991, both in electrical engineering.

She joined the Military Institute of Engineering (IME) of Rio de Janeiro, in 2002, where she is Associate Professor at the Electrical Engineering Department. She founded and heads the Laboratory of Acoustic Signal Processing (LASP). In 2003, she received the University Research Program grant award from CISCO/USA. She also served as editorial board member of the IEEE Communications Surveys and Tutorials from 1999–2007. Since 2008, she has been responsible for the International Scientific Collaboration IME-ParisTech that includes 10 french engineering schools. She was President-Adjoint of the Brazilian Telecommunications Society from 2008–2010 and she is member of the Signal Processing Society. In 2011, Prof. Coelho received the USPTO patent of an automatic speaker recognition method based on a new speech feature and speaker classifier. Her main research interests include acoustic signal processing, speech enhancement and intelligibility, speech and speaker recognition, time-frequency analysis, acoustic emotion detection and classification, acoustic speech features, acoustic signal and noise representation and generation, non-stationary noise, and statistical signal processing.