Colored Noise Based Multicondition Training Technique for Robust Speaker Identification

L. Zão and R. Coelho, *Member, IEEE*

***Abstract*—This letter proposes a colored noise based multicondition training technique for robust speaker identification in unknown noisy environments. The colored noise samples generation is based on filtering a white Gaussian sequence that leads to a power spectral density (PSD) proportional to $1/f^{\beta}$, where $\beta \in [0, 2]$. Gaussian mixture models (GMM) are applied to obtain the speaker models using the noisy speech signals with a single signal-to-noise ratio (SNR). The colored noise based multicondition training is evaluated for the speaker identification task considering the test utterances corrupted with real acoustic noises and different values of SNR. The results show that the proposed technique outperforms the white noise based multicondition and the clean-speech training approaches.**

***Index Terms*—Automatic speaker recognition, colored noises, Gaussian mixture model, multicondition training.**

I. Introduction

In recent years, the improvement of noise robustness in speaker recognition systems became an important issue. The multicondition training technique [1]–[3] was proposed to overcome the degradation of the recognition accuracy in acoustic noisy environments. The idea is to compensate the mismatch between the training and testing phases. The use of artificial noise in multicondition training is also an interesting solution when no information concerning the acoustic noise sources is available [4]. This technique was applied by using white noise [4] with different values of SNR. However, colored spectra have been measured in many environmental acoustic noises [5], [6].

This Letter presents a new approach for multicondition training in automatic speaker recognition applications considering artificial colored acoustic noises. In this proposal, multiple copies of the training utterances are corrupted with noise samples artificially generated with Gaussian pattern and colored spectra. Since it is assumed no knowledge about the real noises, they are not used to corrupt the training speech. The proposed approach differs from the multicondition training in [4] since in this work a single value of SNR is adopted to corrupt the training speech. However, it is not restricted to the single SNR case. The PSD shape of the noise samples is achieved by choosing a filter whose frequency response is proportional to $1/f^{\beta/2}$. This requirement is attained using the Al-Alaoui rule [7] in the filter transfer function. The filter coefficients are calculated via finite-length power series expansion, leading to a FIR (finite impulse response) filter.

Manuscript received July 13, 2011; revised September 02, 2011; accepted September 14, 2011. Date of publication September 26, 2011; date of current version October 10, 2011. This work was supported in part by the Universal/CNPq (472461/2009-5) research grant.

The authors are with the Electrical Engineering Department, Military Institute of Engineering (IME), Rio de Janeiro, Brazil (e-mail: zao@ime.eb.br; 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/LSP.2011.2169453

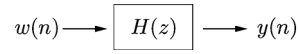

Fig. 1. White Gaussian sequence $w(n)$ is filtered to obtain a colored sequence $y(n)$.

The proposed multicondition training is evaluated for the speaker identification task. GMM models are obtained from the feature vectors extracted from the set of corrupted utterances of each speaker. For the identification tests the speech signals are corrupted by acoustic noises collected from six different sources, considering five values of SNR. The MFCC (mel-frequency cepstral coefficients) speech features and their corresponding dynamic (Δ) coefficients are extracted from the training and testing utterances. The experiments are also conducted with white noise based multicondition and clean-speech training approaches. Additional contributions of this work areas follows.

- The generation of artificial colored noises, with spectra parametrized by $\beta \in [0, 2]$, to represent the real environmental acoustic noises. Three different values of β are adopted in the experiments, without assuming any prior knowledge about the noises sources.
- The definition of a single SNR value to corrupt the training speech utterances. This value is obtained assuming the occurrence of environmental acoustic noises in the tests with SNR between 0 dB and 20 dB.

This letter is organized as follows. Section II describes the colored-spectra noise samples generator. Section III presents the conventional GMM approach and the proposed colored noise based multicondition training technique. Section IV describes the speaker identification experiments conducted with the proposed technique. In the same Section, the identification accuracies are discussed and compared to the baseline approaches. Finally, Section V concludes this Letter.

II. Colored Noise Samples Generation

The PSD of a noise can be represented by its shape $S(f) \propto 1/f^{\beta}$, where β is generally on the range $[0, 2]$. According to the PSD decaying rate, noises can be classified as white ($\beta \approx 0$), pink ($\beta \approx 1$) and brown ($\beta \approx 2$).

The colored noise samples $y(n)$ are obtained by filtering a white Gaussian sequence $w(n)$ (Fig. 1). The PSD of the sequence $y(n)$ is given by $S(f) = \sigma_w^2 |H(e^{j2\pi fT})|^2$, where σ_w^2 is the variance of $w(n)$, and $|H(e^{j2\pi fT})|$ is the filter frequency response.

Considering a finite-length power series expansion, the filter transfer function can be written as

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(N-1)z^{-(N-1)} \quad (1)$$

where $h(k), k = 0, 1, \ldots, N-1$, are the filter coefficients. Each term of the sequence $y(n)$ is thus calculated by the convolution

$$y(n) = w(n) * h(n) = \sum_{k=0}^{N-1} h(k)w(n-k). \quad (2)$$

It follows from (2) that, as $w(n)$, $y(n)$ has also a Gaussian distribution.

The $1/f^{\beta}$ PSD of $y(n)$ is obtained by adopting the Al-Alaoui digital integrator transfer function [7], with $\beta/2$ as the fractional order exponent, to compose the transfer function $H(z)$:

$$H(z) = \left[\frac{7T}{8}\frac{\left(1+\frac{z^{-1}}{7}\right)}{(1-z^{-1})}\right]^{\beta/2} \quad (3)$$

where T is the sampling period.

Thus, the resulting filter frequency response [8] is

$$|H(e^{j2\pi fT})| = \left[\frac{7T}{8}\right]^{\beta/2}\left[\frac{\frac{50}{49}+\frac{2}{7}\cos(\pi fT)}{2\mathrm{sen}(\pi fT)}\right]^{\beta/2}. \quad (4)$$

It can be noted from (4) that the PSD of $y(n)$ follows the relation $S(f) \propto 1/f^{\beta}$, when $f \to 0$.

The filter coefficients are obtained by the convolution $h(k) = a(k) * b(k)$, where $a(k)$ and $b(k)$ are the first $N/2$ terms obtained by expanding, respectively, the numerator and denominator of (3) in power series [8]:

$$a(k) = \left(\frac{-1}{7}\right)\frac{k-\frac{\beta}{2}-1}{k}a(k-1), \quad a(0)=1, \quad (5)$$

$$b(k) = \frac{k+\frac{\beta}{2}-1}{k}b(k-1), \quad b(0)=1. \quad (6)$$

Hence, a noise sequence $y(n)$ is obtained with PSD, or spectrum color, defined by the input β value in (5) and (6).

III. Colored Noise Multicondition Training Proposal

This section presents the colored noise based multicondition training technique for robust speaker identification. Since the GMM models are adopted for the speaker classification, some concepts about the conventional GMM approach [9] are presented before the description of the proposed technique.

A. GMM

The GMM (λ_S) of a speaker S is defined as a linear combination of Gaussian components

$$p(\vec{x}|\lambda_S) = \sum_{j=1}^{M} p_j b_j(\vec{x}) \quad (7)$$

where $\vec{x}$ is a D-dimensional speech feature vector, p_j are the mixture weights, with $\sum_{j=1}^{M} p_j = 1$, and $b_j(\vec{x})$ are the Gaussian densities with mean vectors $\vec{\mu}_j$ and covariance matrices K_j, i.e.,

$$b_j(\vec{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det K_j}} \times \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T K_j^{-1}(\vec{x}-\vec{\mu}_j)\right). \quad (8)$$

Thus, the GMM of speaker S can be parametrized by

$$\lambda_S = \{p_j, \vec{\mu}_j, K_j | j = 1, \ldots, M\}. \quad (9)$$

Let Φ_S^0 denote the training speech utterance of speaker S. The parameters of λ_S are estimated as to maximize the likelihood function

$$p(X|\lambda_S) = \prod_{t=1}^{Q} p(\vec{x}_t|\lambda_S) \quad (10)$$

where the speech feature matrix X is extracted from Φ_S^0, and composed of Q feature vectors $X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_Q]$. The decision rule of the speaker identification task is based on the maximum log-likelihood criteria [10],

$$\hat{S} = \arg\max_{S} \sum_{t=1}^{Q} \log p(\vec{x}_t|\lambda_S) \quad (11)$$

where $\vec{x}_t$ are the test feature vectors and the probability $p(\vec{x}_t|\lambda_S)$ is calculated according to (7). This means that the identified speaker $\hat{S}$ maximizes the sum in (11).

B. GMM in Multicondition Training

Multicondition data sets $(\Phi_S^i, i = 1, 2, \ldots, m)$ are obtained by adding the colored noise sequences to multiple copies of the clean training utterance (Φ_S^0) of each speaker S. The speech feature matrices, extracted from the corrupted data sets, are then used to obtain a set of m GMM models (λ_S^i) for speaker S:

$$p(\vec{x}|\lambda_S^i) = \sum_{j=1}^{M} p_j^i b_j^i(\vec{x}), \qquad i = 1, \ldots, m. \quad (12)$$

According to (12), each speaker model λ_S^i is composed by M Gaussian densities. This leads to a total of $m \times M$ components computed and stored for each speaker. The models of speaker S are parametrized by

$$\lambda_S^i = \{p_j^i, \vec{\mu}_j^i, K_j^i | j = 1, \ldots, M\}, \qquad i = 1, \ldots, m \quad (13)$$

where $\vec{\mu}_j^i$ are the mean vectors and K_j^i are the covariance matrices of the Gaussian densities $b_j^i(x)$. Hence, for speaker S, the colored multicondition training model (Λ_S) is given by the collection of all the parameters estimated in (13),

$$\Lambda_S = \bigcup_{i=1}^{m} \lambda_S^i = \{p_j^i, \vec{\mu}_j^i, K_j^i | i = 1, \ldots, m; j = 1, \ldots, M\}. \quad (14)$$

Considering the speaker models Λ_S, the decision rule adopted in the speaker identification task is given by

$$\hat{S} = \arg\max_{S} \sum_{t=1}^{Q} \log p(\vec{x}_t|\Lambda_S). \quad (15)$$

In this proposal, the probability $p(\vec{x}|\Lambda_S)$ is adjusted to consider all $m \times M$ Gaussian densities for speaker S by the following equation:

$$p(\vec{x}|\Lambda_S) = \sum_{i=1}^{m} \pi_i p(\vec{x}|\lambda_S^i) = \sum_{i=1}^{m}\pi_i \sum_{j=1}^{M} p_j^i b_j^i(\vec{x}) = \sum_{i=1}^{m}\sum_{j=1}^{M} \pi_i p_j^i b_j^i(\vec{x}). \quad (16)$$

Each term π_i in (16) represents the weighting of the noise condition Φ_S^i, with $\sum_{i=1}^{m} \pi_i = 1$. Note that the expression on the right side of (16) is a linear combination of $m \times M$

TABLE I
SPEAKER IDENTIFICATION ACCURACIES (%) WITH THE PROPOSED MULTICONDITION TRAINING

Training SNR	Testing SNR					Average
	0 dB	5 dB	10 dB	15 dB	20 dB	
10 dB	19.90	39.50	56.51	65.90	73.00	50.96
15 dB	17.39	34.57	54.42	66.84	74.28	49.50
20 dB	15.82	28.95	48.60	66.71	78.19	47.65

Gaussian densities $b_j^i(\vec{x})$, with constants coefficients $\pi_i p_j^i$, such that $\sum_{i=1}^{m}\sum_{j=1}^{M}\pi_i p_j^i = 1$. Since no knowledge about noise corruption is assumed, it is adopted $\pi_i = 1/m$ for all $i = 1, \ldots, m$, in the experiments conducted in this work.

IV. EXPERIMENTAL SETUP AND RESULTS

The speaker identification experiments are conducted using the KING speech database to evaluate the proposed colored noise based multicondition training technique. The KING database is composed of ten sessions of speech, spoken by 49 male speakers. The first five sessions are used in the experiments, resulting in 100 s of speech per speaker, in average, after silence extraction. Three of these sessions (60 s of speech) are used for model training. The remaining two sessions are used to evaluate the identification accuracies with 1960 tests of 1 s, and 392 tests of 5 s.

The speech feature vectors are composed by 20 MFCC and their corresponding first order dynamic (Δ) coefficients, extracted from frames of 20 ms with 50% overlapping.

Four environmental acoustic noises (Buccaneer, Destroyer, Factory, and Volvo), and also an artificially generated Gaussian white noise, are collected from NOISEX-92 database [11] to corrupt the test utterances, with SNR varying from 0 to 20 dB. A non-stationary noise (Siren) [12], collected from a fire engine siren, is also included in the experiments.

Three speaker identification experiments are conducted to evaluate the impact of the SNR values adopted in the training utterances corruption. For this purpose, $m = 3$ colored noise sequences are generated as described in Section II, adopting $\beta = 0$ (white noise), $\beta = 1$ (pink noise), and $\beta = 2$ (brown noise). These colors are chosen since such noises occur in many areas of science [5]. In the experiments, the noise sequences are added to the clean training utterances with SNR of 10 dB, 15 dB and 20 dB. Each corrupted speech data set is used to generate a GMM speaker model with $M = 32$ Gaussian densities. Table I presents the identification accuracies obtained in the experiments with tests duration of 5 s. These results are obtained with the 392 test utterances corrupted with the six acoustic noises, leading to 2352 tests and accuracy precision of 0.0425. This value is estimated with a confidence degree of 95% using the Chebyshev inequality [13]. Since the best average identification performance is obtained with SNR of 10 dB, this value is adopted in the other experiments presented in this Letter.

For the colored noise multicondition training evaluation, the speaker identification task is also conducted with the conventional GMM model (Conv-GMM) [10] and the white noise based multicondition training baseline (BSLN-Mul) [4]. In the former, the clean speech signals are used to train GMM models with 32 Gaussian components. In the latter, multiple copies of the training speech signals are corrupted by the Gaussian white noise with SNR varying from 10 to 20 dB, and intervals of 2 dB. The clean and the corrupted speech signals are then concatenated and used to generate the BSLN-Mul models with 128 Gaussian densities. Table II presents the identification accuracies obtained in the experiments with tests duration of 5 s, with accuracies precision of 0.2550.

TABLE II
SPEAKER IDENTIFICATION ACCURACIES (%) FOR TESTS DURATION OF 5 S

Noise	SNR (dB)	Conv-GMM	BSLN-Mul	New Proposal
Clean		91.58	88.01	88.27
Buccaneer	20	58.16	72.96	69.39
	15	38.27	58.67	57.14
	10	20.66	41.33	45.41
	5	9.95	21.17	23.72
	0	2.81	5.36	10.97
Destroyer	20	84.44	83.93	85.46
	15	77.30	75.77	77.55
	10	54.59	54.85	53.83
	5	21.94	25.26	28.06
	0	6.38	10.97	10.71
Factory	20	83.16	85.20	86.73
	15	70.66	80.36	81.63
	10	46.17	67.35	69.90
	5	23.72	35.97	53.57
	0	11.22	5.10	13.52
Siren	20	36.73	30.61	32.65
	15	16.07	13.01	14.03
	10	4.85	6.12	6.12
	5	2.04	2.81	4.34
	0	2.04	2.55	2.81
Volvo	20	90.05	85.46	89.54
	15	87.50	83.16	89.54
	10	82.91	77.81	88.52
	5	71.94	63.78	82.14
	0	47.96	40.56	67.60
White	20	61.73	86.48	74.23
	15	38.01	84.44	75.51
	10	18.88	71.68	75.26
	5	8.16	31.12	45.15
	0	3.06	12.50	13.78
Average		39.38	47.21	50.96

The results show that the proposed technique presents the highest accuracies in comparison to the baseline approaches in 21 experiments, from a total of 30 (6 noises $\times$ 5 SNR). Comparing to the Conv-GMM, the new technique achieves more than 56% of improvement for the white noise and SNR of 10 dB. Considering the real acoustic noises, an increase of 30% is obtained for the Factory noise (SNR of 5 dB).

With respect to the BSLN-Mul, the increase in the identification rates achieves 27% for the Volvo noise and SNR of 0 dB. The proposed approach outperforms the BSLN-Mul for severe noisy conditions ($\mathrm{SNR} \leq 10$ dB) even when the white noise is considered in the tests. This can be explained by the fact that the new proposal adopts a single training SNR of 10 dB, that is more appropriate for this noise levels.

It can be noted that the BSLN-Mul outperforms the new approach for the Buccaneer noise with SNR of 15 dB and 20 dB. However, when these same noise levels are applied in the training phase, the proposed technique achieves superior performances than the BSLN-Mul: 59.18% for SNR of 15 dB and 76.53% for 20 dB. These results show that, for the Buccaneer noise, training with higher noise levels than those in the tests degrades the performance of the proposed approach in comparison to the BSLN-Mul. Considering the Siren noise, the Conv-GMM achieves the highest accuracies in the tests

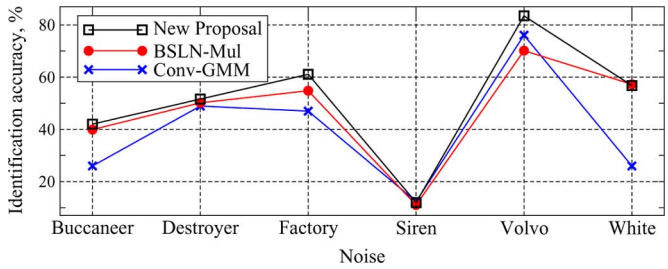

Fig. 2. Average identification accuracies (%) for tests duration of 5 s.

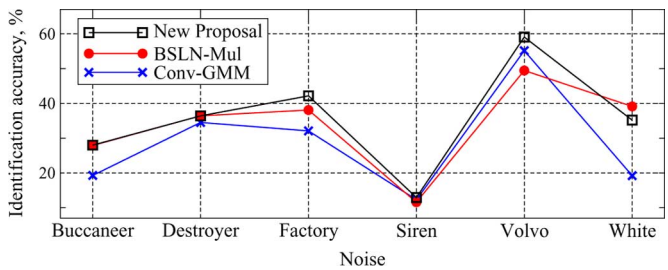

Fig. 3. Average identification accuracies (%) for tests duration of 1 s.

with SNR of 15 dB and 20 dB. However, the new proposal outperforms the BSLN-Mul for all values of SNR, except for the similar identification rates obtained with SNR of 10 dB. The performance obtained for Siren noise can be improved by the use of the post-processing spectral attenuation algorithm proposed for highly non-stationary noises [14].

Tables I and II also demonstrate that other SNR values can be used in the multicondition training. For example, the adoption of SNR of 20 dB leads to an average identification accuracy of 47.65%, while the Conv-GMM and the BSLN-Mul approaches achieve 39.38% and 47.21%, respectively.

Fig. 2 presents a comparison among the proposed and the baseline approaches for the different acoustic noises used to corrupt the test utterances. The curves present the average identification rates obtained considering the five values of SNR shown in Table II. It can be noted that the proposed technique presents superior performances compared to the BSLN-Mul for all the environmental noises. Fig. 3 depicts the same comparison among the average identification accuracies, considering tests duration of 1 s. For this case, the proposed technique also achieves the highest identification rates for the real environmental noises. In comparison to the Conv-GMM, the increase in the average identification accuracy achieves 9.6% for the Volvo noise corruption. For the Factory noise, the proposed approach outperforms the BSLN-Mul with an increase of 10.1% in the average identification rates. Even for the Volvo noise, for which the BSLN-Mul does not improve the Conv-GMM performance, the proposed technique achieves the highest accuracies for both 1 s and 5 s tests durations.

V. Conclusion

This Letter proposed a colored noise based multicondition training technique for robust speaker identification. The main idea is to artificially corrupt the clean-speech training utterances to obtain the GMM speaker models. For this purpose, colored noise sequences were obtained by filtering white Gaussian sequences, resulting in $1/f^{\beta}$ PSD shapes, with $\beta \in [0, 2]$. The description of the noise samples generation method was also presented in this Letter. Speaker identification experiments were conducted to evaluate the proposed multicondition training technique. In the identification tests, the speech signals were corrupted with environmental acoustic noises and different values of SNR. For comparison, the performances of white noise based multicondition and clean-speech training were also examined in the experiments. The results show that the new technique outperforms the white noise based multicondition training approach for all the environmental noises considered in the speaker identification tests.

References

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

[2] D. 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 Language Processing*, 2000, pp. 29–32.

[3] X. Cui and Y. Gong, "Variable parameter gaussian mixture hidden Markov modeling for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003.

[4] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1711–1723, Jul. 2007.

[5] R. Voss and J. Clarke, "$1/f$ noise in music: Music from $1/f$ noise," *J. Acoust. Soc. Amer.*, vol. 63, no. 1, pp. 258–263, 1978.

[6] M. Keshner, "$1/f$ noise," *Proc. IEEE*, vol. 70, pp. 212–218, Mar. 1982.

[7] M. Al-Alaoui, "Novel digital integrator and differentiator," *Electron. Lett.*, vol. 29, pp. 376–378, Feb. 1993.

[8] Y. Ferdi, A. Taleb-Ahmed, and M. Lakehal, "Efficient generation of $1/f^{\beta}$ noise using signal modeling techniques," *IEEE Trans. Circuits Syst.*, vol. 55, pp. 1704–1710, Jul. 2008.

[9] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 639–643, Oct. 1994.

[10] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 72–82, 1995.

[11] 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.

[12] Fire Engine Siren FreeSFX [Online]. Available: http://www.freesfx.co.uk

[13] A. O. Allen, *Probability, Statistics, and Queueing Theory With Computer Science Applications*. Orlando, FL: Academic, 1978.

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