Robust Speaker Verification in Colored Noise Environment

César A. Medina, José A. Apolinário Jr.
IME – DE/3
Praça General Tibúrcio 80
Rio de Janeiro, RJ 22290-270 – Brazil
cancerbero@hotmail.com, apolin@ieee.org

Abraham Alcaim
CETUC/PUC-Rio
Rua Marquês de São Vicente, 225
Rio de Janeiro, RJ 22453-900 – Brazil
alcaim@cetuc.puc-rio.br

Rogerio G. Alves
Clarity Technologies Inc.
3290 W Big Beaver Road S. 220
Troy, MI 48084 – USA
roger@claritytechinc.com

***Abstract*—Noise robustness of automatic speaker verification systems is crucial in real life applications. A study on the performance of several spectral subtraction-based speech enhancement techniques shows the poor performance of these algorithms when used as a preprocessing stage of the speaker verification system in the presence of colored noise. In this paper, a new technique based on the addition of modeled colored noise is introduced. Experimental results in both white and colored noise environments are presented, showing the improvement of the proposed scheme.**

I. Introduction

The effect of additive noise in a speaker verification system is well known to be a crucial problem in real life applications. Speech enhancement algorithms, such as the spectral subtraction based techniques, have been widely used to reduce the effect of additive noise in several areas of speech processing [1]-[3].

In the speaker verification problem, where a binary decision is required—accepting or rejecting the pretense speaker—, if the test utterance is corrupted by any type of noise, the performance of the system notoriously degrades. This is true even if the speech signals are preprocessed with spectral subtraction techniques. Simulation results with different types of noise are given in this paper.

Aiming at improving the robustness of the speaker verification system with spectral subtraction techniques, we introduce in this paper the use of modeled colored noise. The purpose is to obtain a noise magnitude spectrum more similar to the one present in the test signal. This will allow a more efficient training of the speaker verification system and, therefore, an improved performance.

The paper is organized as follows. Section II reviews the basic concepts of classical speech enhancement schemes. In Section III, we describe the speaker verification system use throughout this work. In Section IV, we present the simulations results in noisy environments with and without spectral subtraction-based speech enhancement. Then, a new scheme for speaker verification in colored noisy environments is introduced and analyzed in Section V. Finally, Section VI summarizes the main conclusions of this work. .

II. Spectral Subtraction–Based Speech Enhancement Methods

Boll [1] has carried out the first detailed investigation on this type of algorithms which try to recover a signal $s = \{s_i\}$ from observations of a noisy signal $d_i = s_i + n_i$, $i = 1, \cdots, N$, where $\{n_i\}$ are independent and identically distributed Gaussian variables with zero mean and variance $\sigma^2_{n_i}$.

Assuming that the human ear is not very sensitive to the phase of a signal, the estimate of s, $\hat{s}$, can be calculated as $\hat{s} = FFT^{-1}\left\{|\widehat{S}(w)|e^{j\phi_{\mathcal{D}}(w)}\right\}$, where $\phi_{\mathcal{D}}(w)$ is the phase of the noisy signal and $|\widehat{S}(w)|$ is the estimate of the absolute value of the Fast Fourier Transform (FFT) of the clean signal computed as $|\widehat{S}(w)| = G(w) \cdot |\mathcal{D}(w)|$, $\mathcal{D}(w)$ being the FFT of the noisy signal. There are several ways to obtain $G(w)$. We have used three methods, which are briefly described as follows.

A. Power Spectral Subtraction

The first step of this algorithm, as in all other spectral subtraction based algorithms, corresponds to the estimation of the noise present in the noisy speech signal. The noise estimate is obtained from the silence frames of the speech signal and is computed as follows:

$$|\widehat{\mathcal{N}}(w)|^2 = \lambda|\widehat{\mathcal{N}}(w)|^2 + (1-\lambda)|\mathcal{D}(w)|^2 \qquad (1)$$

where λ is the *forgetting factor* and determines a trade-off between the variance of the estimated spectrum and the ability of tracking fast time variations on the statistics of the noise.

The function $G(w)$ is given by the following expression:

$$G(w) = \begin{cases} \left(1 - \frac{|\widehat{\mathcal{N}}(w)|^2}{|\mathcal{D}(w)|^2}\right)^{1/2}, & |\widehat{\mathcal{N}}(w)|^2 \leq |\mathcal{D}(w)|^2 \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

B. Ephraim–Malah Filter

Ephraim and Malah proposed in [2] an estimate $|\widehat{S}(w)|$ such that the so-called *musical artifact* (distortion typically caused by power spectral subtraction) is slightly diminished. This algorithm can be summarized as follows:

0-7803-8104-1/03/$17.00 ©2003 IEEE

- compute the *a posteriori* SNR (Signal-to-Noise-Ratio) and the *a priori* SNR in frame n:

$$\begin{aligned} \gamma_k(n) &= \frac{|\mathcal{D}_k(n)|^2}{|\widehat{\mathcal{N}}_k(n)|^2} \\ \xi_k(n) &= \alpha G^2(\gamma_k(n-1))\gamma_k(n-1) \\ &\quad +(1-\alpha)P[\gamma_k(n)-1] \end{aligned} \tag{3}$$

  where $G(\gamma_k(n)) = \sqrt{1-1/\gamma_k(n)}\ P[\gamma_k(n) - 1]$, $|\widehat{\mathcal{N}}_k(n)|^2$ is computed as in (1) and $P[\cdot]$ is used to guarantee that $\xi_k(n)$ is always positive (it is defined as x if $x \geq 0$ and 0 otherwise);
- define q_k as the probability of absence of the speech signal in the spectral component k. In order to simplify the notation, we also define $\mu_k = (1-q_k)/q_k$, $\eta_k = \xi_k/(1-q_k)$, and $\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k$;
- with these variables, the generalized likelihood ratio is expressed as $\Lambda(\eta_k, \gamma_k, q_k) = \mu_k \frac{e^{\nu_k}}{1+\eta_k}$;
- finally, the filter function, G, is expressed as

$$G(\xi_k, \gamma_k, q_k) = \frac{\Lambda(\eta_k, \gamma_k, q_k)}{1+\Lambda(\eta_k, \gamma_k, q_k)} G_{MMSE}(\xi_k, \gamma_k) \tag{4}$$

  where

$$G_{MMSE}(\xi_k, \gamma_k) = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} e^{\frac{-\nu_k}{2}} \cdot \cdot \left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] \tag{5}$$

  and $\Gamma(\cdot)$ is the Gamma function, $\Gamma(1.5) = \sqrt{\pi}/2$ and $I_0(\cdot)$ and $I_1(\cdot)$ are the zero and first orders modified Bessel functions, respectively. In the simulations carried out in this work, we have used $\alpha = 0.99$ and $q_k = 0.2$ [2].

C. Virag's Method

Virag [3] has proposed a method that uses a generalized spectral subtraction function as given by

$$G(w) = \begin{cases} \left(1 - \alpha\left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma} < \frac{1}{\alpha+\beta} \\ \left(\beta\left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \text{otherwise} \end{cases} \tag{6}$$

where α, typically between 1 and 6, is the over–subtraction factor that decreases musical noise but increases signal distortion, β (usually $0 \leq \beta \ll 1$) is the spectral floor that decreases musical noise but increases background noise.

The method introduced by Virag uses a masking threshold, $T(k)$, to adapt optimally (in the sense of hearing perception) the coefficients α and β. The value of γ is fixed to 2. The adaptation is carried out, for each coefficient of each segment of speech being analyzed, by means of a linear interpolation $\alpha_k = F_\alpha[\alpha_{min}, \alpha_{max}, T(k)]$ and $\beta_k = F_\beta[\beta_{min}, \beta_{max}, T(k)]$ where α_{min}, α_{max}, β_{min} and β_{max} are the maximal and the minimal values of the over–subtraction and spectral floor coefficients, $\alpha_{min} = 1$, $\alpha_{max} = 6$ and $\beta_{min} = 0$, $\beta_{max} = 0,01$. F_α and F_β are linear interpolation functions such that $F_\alpha = \alpha_{max}$ if $T(k) = T(k)_{min}$ and $F_\alpha = \alpha_{min}$ if $T(k) = T(k)_{max}$, where $T(k)_{min}$ and $T(k)_{max}$ are the minimal and the maximal values of $T(k)$ for the speech segment being analyzed. Similar reasoning is used to calculate F_β.

The computation of the masking threshold, $T(k)$, is carried out for each interval of speech under analysis and is based on a hearing perception model as in [4].

III. The Automatic Speaker Verification System

Our experiments were carried out on a speaker verification system implemented with 15 mel-cepstral coefficients and the Gaussian Mixture Model (GMM) [5]. The input signal features and the previously stored model of the pretense speaker are used to decide for *acceptance* or *rejection* of that speaker.

The mixture of Gaussian probability densities is a weighted sum of M densities given by

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x}) \tag{7}$$

where $\boldsymbol{x}$ is a random vector of dimension N, $b_i(\boldsymbol{x})$, $i = 1, \ldots, M$, are the densities and p_i, $i = 1, \ldots M$, are the weights of the mixture. Each density is an N-dimensional Gaussian function of the form

$$b_i(\boldsymbol{x}) = \frac{e^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{K}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)}}{(2\pi)^{\frac{N}{2}}\sqrt{|\boldsymbol{K}_i|}} \tag{8}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{K}_i$. The weights are such that condition $\sum_{i=1}^{M} p_i = 1$ is satisfied.

The Gaussian mixture densities are parameterized by a mean vector, a covariance matrix, and the weighting of the mixture components (λ model). These parameters are jointly represented by $\lambda = \{p_i, \boldsymbol{\mu}_i, \boldsymbol{K}_i\}$, $i = 1, \ldots, M$.

The speaker verification system must decide if a speech utterance $\boldsymbol{X}$ belongs (or not) to a given speaker with a previously obtained λ model. The two possible hypotheses are:

H_0: $\boldsymbol{X}$ belongs to the speaker.

H_1: $\boldsymbol{X}$ does not belong to the speaker.

In the specification of the likelihood test (to decide between H_0 and H_1), we normally use a model for a universe of false probabilities, namely, the *background* model. It is built from a set of false speakers representing possible impostors to the system. In the logarithmic domain, the likelihood ratio is given by $\Lambda(\boldsymbol{X}) = \log p(\boldsymbol{X}|\lambda_L) - \log p(\boldsymbol{X}|\lambda_B)$ where, λ_L is the model of the alleged speaker and λ_B is the *background* model. If this likelihood is greater than a previously defined threshold, the speaker is accepted, otherwise, he is rejected or classified as an impostor. The likelihood for a true speaker is directly computed via

$$\log p(\boldsymbol{X}|\lambda_L) = \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_L) \tag{9}$$

Note that a scale factor $\frac{1}{T}$ was used to normalize the likelihood according to the duration of the utterance (number T of feature vectors).

IV. ASV Performance in Noisy Environment

In all computer simulations carried out with the automatic speaker verification system described in the previous section, two speech data bases were used. The first one, sampled at $8kHz$, with the training utterances corresponding to 60 male speakers, 10 of them used to form the background. Training utterances consisted of 120 seconds of clean speech and silence. Testing utterances were 25 seconds speech signals also including periods of silence. The data base was divided such that a total of 23400 false tests and 600 true tests were performed. Each speech segment of $32ms$, with 50% overlapping, was multiplied by a Hamming window and passed through a pre–emphasis filter $(1 - 0.95z^{-1})$. The second data base used to corrupt the clean signals was NOISEX-92 that contains samples of different types of noise from which we have used: factory noise, aircraft cockpit noise, and *speech like* noise. We have also used artificially produced white Gaussian noise.

The speaker verification system uses 15 mel–cepstral features and a 32 Gaussian Mixture Model (GMM) with universal background model. The performance of the system was carried out with clean training signals and noisy testing signals with SNR ($10log\left(\sigma_s^2/\sigma_n^2\right)$, σ_s^2 being the variance of clean testing signal and σ_n^2 the noise variance such that the variance of the testing signal corresponds to $\sigma_d^2 = \sigma_s^2 + \sigma_n^2$) values ranging from -5 to $10dB$. Tab. I shows the results obtained in terms of Equal Error Rate (EER)—threshold set such that the *false alarm* error rate was identical to the *miss* error rate [6]—for each of the speech enhancement methods used in the preprocessing stage (power spectral subtraction or SS, Ephraim–Malah filter or EMF, and Virag's Method) as well as for the case of no preprocessing (WithOut Speech Enhancement or WOSE). As seen in Tab. I, the error obtained from the automatic speaker verification system is too high for all types of noise and all spectral subtraction-based speech enhancement schemes. This means that even with the speech enhancement techniques being used as a preprocessing stage, the system can not be employed in noisy environments. The strategy presented in the next section aims at obtaining a more robust scheme in these situations.

TABLE I

EER (%) FOR CLEAN TRAINING SIGNALS AND TESTING SIGNAL CORRUPTED WITH ADDITIVE NOISE AT DIFFERENT SNR.

SNR	WOSE	SS	EMF	Virag
White noise				
−5	47.58	47.76	45.76	**44.26**
0	44.76	47.59	45.92	**37.60**
5	41.10	49.09	44.43	**31.28**
10	28.79	47.92	41.60	**24.00**
Speech Like noise				
−5	29.95	49.92	43.10	**23.96**
0	20.30	45.59	38.94	**15.97**
5	**8.99**	46.08	35.44	12.31
10	**4.66**	43.26	30.78	9.65
Cockpit noise				
−5	49.25	48.09	**46.59**	47.25
0	46.92	47.75	41.93	**41.43**
5	43.10	46.92	39.60	**38.44**
10	33.28	46.26	33.28	**28.29**
Factory noise				
−5	44.59	48.42	44.09	**36.44**
0	37.60	48.59	40.43	**24.63**
5	25.29	47.25	38.10	**14.48**
10	11.15	44.93	36.11	**10.98**

V. Proposed Scheme

The proposed method is depicted in Fig. 1 and explained as follows.

After a VAD (Voice Activity Detection) algorithm, from the speech segments already classified as speech or silence, all "silence" segments are extracted. With these segments (noise only), noise's power and magnitude spectrum are estimated. The magnitude spectrum may then be modeled as an AR (Auto-Regressive), MA (Moving-Average), or ARMA (Auto Regressive Moving Average) process. We have chosen to use a linear prediction coefficients (LPC) model and minimize, during analysis, a forward prediction mean square error such that the resulting synthesis model is defined by [7]:

$$H_{LPC} = \frac{1}{1 + \sum_{i=1}^{N_{LPC}} a_{LPC}(i) z^{-i}} \tag{10}$$

where a_{LPC} are the LP coefficients. For the computation of these coefficients, an efficient algorithm like the Levinson–Durbin recursion [7] can be used.

In the simulations, we have used 15 LP coefficients to model the filter H_{LPC}. From a white noise input, this filter is expected to generate a noisy output with approximately the same magnitude spectrum of the noise present in the testing speech signal.

When applying speech enhancement to the testing signal, we estimate its variance (σ_d^2) such that, along with the estimate of the noise variance (σ_n^2), we can estimate its SNR and use this information to obtain the multiplying constant $gain(\sigma_n)$ (see Fig. 1) responsible for matching the SNR of the training signal $(\sigma_t^2/gain^2(\sigma_n))$ with the SNR of the testing signal $(\sigma_s^2/\sigma_n^2)^1$.

$$gain(\sigma_n) = \sqrt{\frac{\sigma_t^2}{\sigma_d^2 - \sigma_n^2}}\sigma_n \tag{11}$$

where σ_t^2 corresponds to the variance of the (clean) training signal.

The following step consists in adding the modeled noise to the clean training signal. The speaker verification system is now trained with the corrupted signal. For both training and testing, a speech enhancement scheme is applied in the preprocessing stage, i.e., just before the speaker verification procedure.

Results obtained from this method are presented in Tab. II. From this table, it can be observed that the spectral subtraction and Ephraim–Malah filter show a great improvement in the performance, but the error rates are higher than those obtained

[1] Note that when $\sigma_s^2 = \sigma_t^2$ (both clean training and clean testing signals having the same power), the gain constant is equal to σ_n.

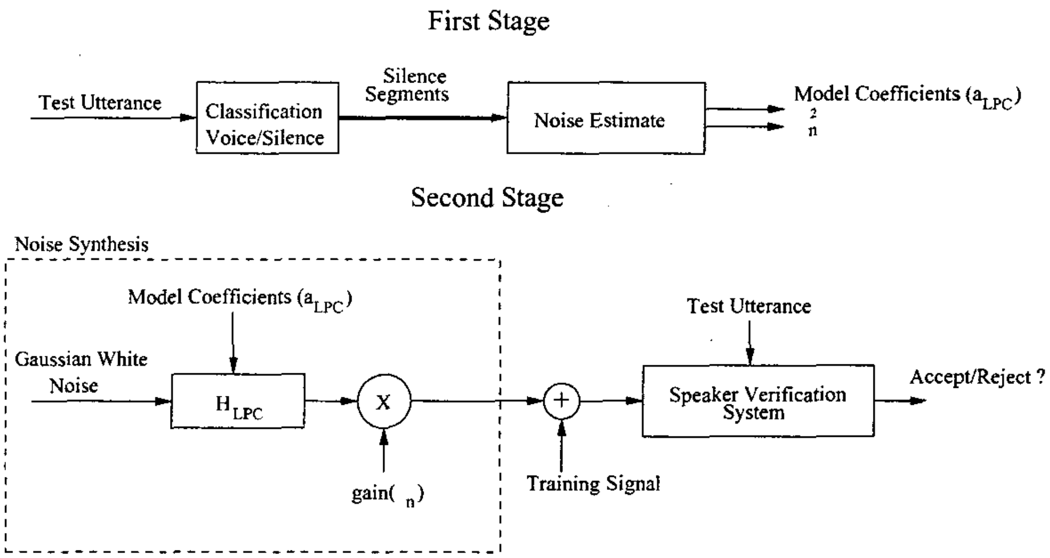

Fig. 1. Speaker Verification (SV) with Modeled Noise Insertion: please note that the SV System performs a speech enhancement in each input signal (noisy *Test Utterance* and artificially corrupted *Training Signal)* before the verification process.

TABLE II

EER (%) FOR TRAINING SIGNAL CORRUPTED BY ADDITIVE MODELLED NOISE.

SNR	SS	EMF	Virag
White Noise			
−5	21.46	10.98	**9.65**
0	19.47	10.15	**7.15**
5	16.64	7.82	**7.15**
10	9.98	6.99	**4.66**
Speech Like Noise			
−5	20.80	5.66	**4.16**
0	8.82	3.33	**2.50**
5	6.16	2.33	**1.16**
10	4.16	2.00	**1.16**
Airplane Cockpit Noise			
−5	25.79	10.98	**7.99**
0	12.98	6.32	**4.83**
5	7.49	4.00	**3.33**
10	6.16	3.00	**2.50**
Factory Noise			
−5	40.60	36.61	**16.97**
0	20.63	16.97	**6.65**
5	9.15	7.49	**4.16**
10	7.32	4.16	**3.16**

with Virag's method. The promising results (significantly lower error rates when compared to those of Tab. I) achieved by Virag's method with the proposed scheme, make a speaker verification system using this approach a good option when operating in presence of colored noise environments.

VI. Conclusions

In this paper, we propose an efficient strategy of speech enhancement for automatic speaker verification (ASV) applications. A linear prediction model of the noise signal and a clean training signal are used to train the ASV system. A spectral subtraction (SS)-based speech enhancement scheme is employed in the preprocessing stage of both the training and test procedures. We have considered three SS methods, and four types of additive noise at different SNR values. The results have shown that the proposed approach significantly improves the ASV robustness in noisy environments, and represent a promising strategy for such applications.

References

[1] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, Apr. 1979, pp. 200–203.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[4] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[5] D. A. Reynolds, "A Gaussian mixture modeling approach to text independent speaker identification," Ph.D. dissertation, Georgia Institute of Technology, 1992.

[6] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M.Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings* of the European Conference on Speech Technology", pp. 1895–1989, 1997.

[7] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, Sept. 1991.