International Conference on Signals and Electronic Systems
September 13-15, 2004 Poznań, POLAND
ICSES'04

Improving the Performance of Speaker Verification Systems in Noisy Environments*

C. A. Medina#, J. A. Apolinário Jr.#, and A. Alcaim##

#IME - Department of Electrical Engineering
Praça General Tibúrcio, 80 – 22290-270, Rio de Janeiro, RJ, Brazil
e-mail: cancerbero@hotmail.com, apolin@ieee.org

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

***Abstract* – Classical speech enhancement techniques and recently developed wavelet denoising schemes are applied to speaker verification systems in noise. Merely applying these techniques to corrupted testing signals does not properly decreases the error rates when clean speech is used for training signals. In this paper, a noise modelling approach is used to corrupt the training signals according to an estimate of the noise present in the test signal. We show that this procedure makes the error rates drop to a fraction of the original results.**

Keywords: speech enhancement, speaker verification, wavelet denoising.

I. Introduction

In view of its several important applications, such as forensic speaker verification, automatic speaker recognition has been the focus of intensive research. Among the problems related to the low performance of these systems in practical implementations, the operation in noisy environments plays an important role due to its devastating effect as SNR decreases. In this work, we investigate how effectively the classical and modern speech enhancement techniques can mitigate the effects of additive white and colored noise. Moreover, assuming that we have clean training signals and noisy testing signals, we show how to further improve the results when noise estimation and modelling is used as an attempt to impose the same noisy environment in both (training and testing) signals.

This paper is organized as follows. Section II reviews the fundamentals of spectral subtraction methods used for speech enhancement. In Section III, the most recently proposed wavelet denoising schemes applied to speech enhancement are discussed. Then, a short description of the speaker verification system is presented in Section IV. In the next section, speech enhancement techniques are applied to the speaker verification system and the noise modelling approach is used to refine the results. The main conclusions are summarized in Section VI.

*Authors thank FAPERJ, CNPq, and CAPES for partical funding of this paper.

II. Spectral subtraction-based speech enhancement

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 σ_i^2. The enhancement filter $G(w)$ is used to provide an estimate of the short-time amplitude spectrum (STAS) of the clean signal as $|\widehat{\mathcal{S}}(w)| = G(w) \cdot |\mathcal{D}(w)|$, $\mathcal{D}(w)$ being the FFT of the noisy signal. There are several ways to obtain $G(w)$. We have considered three methods, which are briefly described as follows.

A. Power Spectral Subtraction

The first step of this algorithm corresponds to the estimation of the noise present in the speech signal. The noise estimate is obtained from the silence frames of the speech signal and is computed as $|\widehat{\mathcal{N}}(w)|^2 = \lambda|\widehat{\mathcal{N}}(w)|^2 + (1-\lambda)|\mathcal{D}(w)|^2$, 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:

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

B. Ephraim–Malah Filter

Ephraim-Malah algorithm [2] uses a minimum mean-square error STAS estimator. The *a posteriori* and *a priori* SNR's in frame n are computed as:

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

where $G(\gamma_k(n)) = \sqrt{1 - 1/\gamma_k(n)}P[\gamma_k(n)-1]$, $|\widehat{\mathcal{N}}_k(n)|^2$ is obtained as in the previous scheme, 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). The filter function, G, is expressed as

$$G(\xi_k, \gamma_k, q_k) = \frac{\Lambda(\eta_k, \gamma_k, q_k)}{1 + \Lambda(\eta_k, \gamma_k, q_k)} G_{MMSE}(\xi_k, \gamma_k) \tag{3}$$

where

$$\begin{aligned} G_{MMSE}(\xi_k, \gamma_k) &= \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} e^{\frac{-\nu_k}{2}} \\ &\times \left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right], \end{aligned} \tag{4}$$

q_k is the probability of absence of the speech signal in the spectral component k, $\mu_k = (1-q_k)/q_k$, $\eta_k = \xi_k/(1-q_k)$, $\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k$, $\Lambda(\eta_k, \gamma_k, q_k) = \mu_k \frac{e^{\nu_k}}{1+\eta_k}$, $\Gamma(\cdot)$ 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

$$G(w) = \begin{cases} \left(1 - \alpha\left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma} < \frac{1}{\alpha+\beta} \\ \left(\beta\left[\frac{|\hat{\mathcal{N}}(w)|}{|\mathcal{D}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \text{otherwise} \end{cases} \tag{5}$$

where α, typically between 1 and 6, is the over–subtraction factor that decreases *musical artifact* but increases audible distortion, β (usually $0 \leq \beta \ll 1$) is the spectral floor that decreases musical noise but increases background noise. A signal masking threshold, $T(k)$, is used to optimally adapt (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 $\alpha_{min} = 1$, $\alpha_{max} = 6$ and $\beta_{min} = 0$, $\beta_{max} = 0,01$. 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. WAVELET-BASED SPEECH ENHANCEMENT

Wavelet-based speech enhancement [5]-[7] can be formulated in three steps. Firstly, a wavelet decomposition of J levels is applied to the input signal. Then, a nonlinear (hard or soft) thresholding function is applied to the detail coefficients of the transform. The soft–thresholding function, used in this work, is defined as:

$$\hat{\alpha}_{jk} = \begin{cases} \text{sgn}(\beta_{jk})(|\beta_{jk}| - t), & |\beta_{jk}| \geq t \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where β_{jk} is the k-th detail coefficient of the noisy signal in level j, $j = 1, \cdots, J$, $k = 1, \cdots, n_j$, and $n_j = \frac{N}{2^{J-j+1}}$. Finally, the wavelet transform is inverted to obtain the enhanced signal $\hat{s} = \hat{s}_i$. In [5], it is claimed that the *VisuShrink* technique, when used with a soft–thresholding, achieves near noiseless reconstructions. Donoho and Johnstone [5] proposed an estimate of the universal thresholding from $\hat{t} = \hat{\sigma}\sqrt{2\log N}$, where N is the total number of coefficients and $\hat{\sigma}$ is an estimate of the noise level given by $\hat{\sigma} = m/0.6745$, m being the median absolute deviation (MAD) of the highest ($j = J$) resolution level detail coefficients. In this paper, we use level dependent thresholding where m is replaced by m_j — the MAD at level j of the transform detail coefficients.

The *SureShrink* algorithm [5]-[7] uses the Stein's method of unbiased risk estimation (SURE). The k-th detail coefficient of the wavelet transform of the noisy signal at level j is defined by $\beta_{jk} = \alpha_{jk} + \sigma_j \xi_{jk}$, where $k = 1, \cdots, n_j$, $\sigma_j = m_j/0.6745$ are estimates of the noise level and ξ_{jk} are independent random Gaussian variables with zero mean and unit variance. The mean square risk of $\hat{\alpha}$ at level j is defined as $\mathcal{R}_j(\hat{s}, s) = \sum_{k=1}^{n_j} E[(\hat{\alpha}_{jk} - \alpha_{jk})^2]$. The threshold t_j, at level j, must be chosen so as to minimize $\mathcal{R}_j(\hat{s}, s)$. In practice, we do not know α_{jk}. However, we can choose a t_j that minimizes an unbiased risk estimator, $\widehat{\mathcal{R}}_j(\hat{s}, s)$, which is a function of σ_j and β_{jk}, $k = 1, \cdots, n_j$. It can be shown that, for soft–thresholding, the minimization of $\widehat{\mathcal{R}}_j(\hat{s}, s)$ leads to

$$\hat{t}_j = \min_{t \geq 0} \sum_{k=1}^{n_j} (2\sigma_j^2 + t^2 - \beta_{jk}^2) I\{|\beta_{jk}| \geq t\} \tag{7}$$

where $I(x)$ is such that if x is a logical variable assuming the values of *true* or *false*, then $I(x) = 1$ if x is *true*, and $I(x) = 0$ otherwise.

The use of a neural network to obtain the threshold was recently proposed in [8]. This method uses a *back-propagation* neural network with one hidden layer. The net is designed such that its output is an estimate $\hat{t}_j$ that minimizes the mean square error between this estimate and the ideal threshold, i.e., the one that minimizes $\mathcal{R}_j(\hat{s}, s)$. The inputs of the neural network are the MAD and the variance of the detail coefficients at each level of the wavelet decomposition.

IV. 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) [9]. The input signal features and the 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 $p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x})$, 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 with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{K}_i$. The Gaussian mixture densities of the λ model are parameterized 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. In the specification of the likelihood test, 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 $\log p(\boldsymbol{X}|\lambda_L) = \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_L)$. Note that a scale factor $\frac{1}{T}$ was used to normalize the likelihood according to the duration of the utterance (T is the number of feature vectors).

V. IMPROVING THE RESULTS

In the simulations, we have used two data bases. The first one with speech utterances for testing and training, consisting of 60 native Brazilian Portuguese male speakers, 10 of them used to form the *background*. The sample frequency was set to $8\ kHz$. The utterances were used to form a training base with 2 minutes signals, and a test base with 25 seconds signals, a total of $23,400$ false tests and 600 true tests. Each signal was segmented into 32 ms with 50% of overlapping and each segment was multiplied by a Hamming window. A pre–emphasis filter ($1 - 0.95z^{-1}$) was applied. The second data base used to corrupt the clean signals was NOISEX-92. It 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.

In the first experiment, the training signals were noiseless and the testing signals were corrupted with white noise at different SNR levels. The testing signals were pre-processed with six different speech enhancement algorithms: power spectral subtraction (SS), Ephraim-Malah filter (EMF), Virag's method, *SureShrink* and *VisuShrink* wavelet–based techniques, and the neural network–based scheme introduced in [8] (NN). Tab. I shows the results in terms of Equal Error Rate (EER) for each case, including the case of no pre-processing or WithOut Speech Enhancement (WOSE). It can be seen that Virag's method yields the best results. Nevertheless, the performance is very poor for all cases.

TABLE I

EER (IN %) FOR CLEAN TRAINING SIGNALS AND CORRUPTED (WHITE NOISE) TESTING SIGNALS.

SNR	WOSE	SS	EMF	Virag	*Sure*	*Visu*	NN
-5	47.6	47.8	45.8	**44.3**	47.9	50.6	49.8
0	44.8	47.6	45.9	**37.6**	47.6	45.9	48.4
5	41.1	49.1	44.4	**31.3**	44.9	43.9	45.8
10	28.8	47.9	41.6	**24.0**	43.1	39.6	39.6

The second experiment was also carried with white noise but the training signal was corrupted with white noise such that its SNR was $5\ dB$. The same speech enhancement algorithms were used in the pre-processing stage and the new values of EER are shown in Tab. II. Comparing to Tab. 1, the results are significantly improved. The good results of this table, when the SNR of the training signal matches the SNR of the testing signal, suggest, as expected, that the training signal should be corrupted with exactly the same amount of noise (same SNR). On the other hand, additional simulations have shown that these improved results–due to adding white noise to the training signal–did not occur when colored noise was present in the testing signal.

TABLE II

EER (%) FOR THE CASE OF BOTH TESTING (DIFFERENT VALUES OF SNR) AND TRAINING (SNR=5 dB) SIGNALS CORRUPTED WITH WHITE NOISE.

SNR	SS	EMF	Virag	*Sure*	*Visu*	NN
-5	36.3	**33.6**	35.9	34.6	33.8	**33.6**
0	24.5	18.3	17.8	**16.3**	16.5	**16.3**
5	16.0	8.6	6.5	**6.3**	**6.3**	**6.3**
10	13.8	8.2	8.3	8.2	**7.8**	8.0

In [10], an alternative approach was suggested when colored noise was corrupting the testing signals. This new scheme is based on modelling the noise embedded in the testing signal. LPC coefficients as well as power level were estimated from those frames of the testing signal where only noise was present. From these estimates, it is possible to synthesize colored noise with a power spectrum density that approximates the one of the noise present in the testing signal. The resulting modelled noise at a proper level is added to the training signal. This allows the verification process to be carried out in similar conditions for training and testing. The results of this method, for three different types of noise and different values of SNR, are shown in Tab. III. It can be seen that now the performance improvements are significant not only for different types of noise, but also at different SNR values.

TABLE III

EER (%) FOR TRAINING SIGNALS CORRUPTED BY MODELLED COLORED NOISE ACCORDING TO [10].

SNR	SS	EMF	Virag	*Sure*	*Visu*	NN
White Noise						
−5	21.5	11.0	9.6	8.6	**8.3**	8.6
0	19.5	10.2	7.2	7.2	7.2	**6.7**
5	16.6	7.8	7.2	**6.0**	6.2	6.2
10	10.0	7.0	**4.7**	5.2	4.8	5.3
Speech Like Noise						
−5	20.8	5.7	**4.2**	5.2	5.2	4.3
0	8.8	3.3	**2.5**	2.8	2.8	2.7
5	6.2	2.3	**1.2**	2.3	2.3	2.2
10	4.2	2.0	**1.2**	2.0	2.0	2.0
Aircraft Cockpit Noise						
−5	25.8	11.0	**8.0**	11.6	11.5	10.8
0	13.0	6.3	**4.8**	5.3	5.2	**4.8**
5	7.5	4.0	3.3	3.3	3.3	**3.2**
10	6.2	3.0	**2.5**	2.7	2.7	2.8
Factory Noise						
−5	40.6	36.6	17.0	16.3	16.3	**16.0**
0	20.6	17.0	6.6	6.2	**5.7**	**5.7**
5	9.2	7.5	4.2	**3.0**	**3.0**	3.2
10	7.3	4.2	3.2	**2.8**	**2.8**	**2.8**

VI. CONCLUSIONS

From the results presented in this paper, a number of important and interesting conclusions can be drawn. When the testing signal is corrupted by noise, the use of clean training signal is definitely not appropriate, independently of the speech enhancement strategy. In case of using the novel noise modelling approach, Virag's method results in the best performance when compared to other spectral subtraction schemes. When both the wavelet-based neural network speech denoising and the noise modelling are used, the results obtained are, in more than 50% of the cases, superior to those obtained by the other two wavelet-based methods. In 50% of the cases, the use of the recently proposed neural network wavelet denoising outperforms Virag's method when both use noise modelling.

Nevertheless, if a confidence interval is taken into account, we could say that the performances of both methods are statistically equivalent. The improved error rates yielded by the noise modelling scheme is a compelling motivation for using this approach in practical implementations where different types of additive colored noise may be present.

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. L. Donoho and I. M. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," in *Proceedings of the 16th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, Maryland, USA, Nov. 1994, pp. 24a–25a.

[6] D. L. Donoho, "De–noising by soft–thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[7] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", *Journal of the American Statistical Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[8] C. A Medina, A. Alcaim, and J. A. Apolinário Jr., "Wavelet denoising of speech using neural networks for threshold estimation," *Electronics Letters*, vol. 39, no. 25, pp. 1869-1870, Dec. 2003.

[9] D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, 1992.

[10] C. A Medina, J. A. Apolinário Jr., A. Alcaim, and R. G. Alves, "Robust speaker verification in colored noise environment," in *Proceedings of the Thirty-Seventh Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, USA, Nov. 2003.