OVERVIEW OF SPEECH ENHANCEMENT TECHNIQUES FOR AUTOMATIC SPEAKER RECOGNITION

Javier Ortega-García and Joaquín González-Rodríguez

Dept. de Ingeniería Audiovisual y Comunicaciones
Universidad Politécnica de Madrid
Crta. Valencia km. 7, Campus Sur, E-28031 Madrid, Spain
e-mail: jortega@diac.upm.es

ABSTRACT[*]

Real world conditions differ from ideal or laboratory conditions, causing mismatch between training and testing phases, and consequently, inducing performance degradation in automatic speaker recognition systems [1]. Many strategies have been adopted to cope with acoustical degradation; in some applications of speaker identification systems a clean sample of speech, prior to the recognition stage, is needed. This has justified the use of procedures that may reduce the impact of acoustical noise on the desired signal, giving rise to techniques involved in the enhancement of noisy speech [2, 9].

In this paper, a comparative performance analysis of single-channel (based in classical spectral subtraction and some derived alternatives), dual-channel (based in adaptive noise cancelling) and multi-channel (using microphone arrays) speech enhancement techniques, with different types of noise at different SNRs, as a pre-processing stage to an ergodic HMM-based speaker recognizer, is presented.

1. INTRODUCTION

Speaker Identification is becoming a high-relevant task in many fields, specially in the framework of security remote applications. These systems, usually developed under laboratory conditions, severely degrade their performance level when an acoustical mismatch appears among training and testing phases. This problem has limited the development of real-world non-specific applications, as testing conditions are higly variant or even unpredictable during the training process.

This mismatch problem has guided to design robust speaker recognizers. The process of providing robustness to the recognizer can be accomplished in three different stages: *i*) the acoustical stage, giving rise to speech enhancement techniques that may improve the SNR of the input signal, *ii*) the parametric stage, by means of parametric representations of speech characteristics which may show inmunity to the noise process and *iii*) the modeling stage, combining adecuate models of noise and clean signal in order to recognize noisy speech.

[*] This work has been supported by CICYT under Project TIC94-0030.

In this paper, a wide analysis of techniques providing robustness to a speaker identification system in the acoustical stage is presented. Section 2 describes single-channel alternatives to speech enhancement, based in the well-known spectral subtraction procedure [3]. In order to solve the problem derived from the appearance of "musical noise", two other alternative techniques are used: spectral subtraction with oversubtraction model [4] and non-linear spectral subtraction [5]. Section 3 faces the problem of multi-channel spech enhancement, providing, on the one hand, a dual-channel optimal solution based on adaptive noise cancellation [6], and on the other hand, a multisensor array performing delay-and-sum beamforming [7]. Section 4 describes the database and the identification system used and shows how this system works when the enhancement algorithms described in sections 2 and 3 are applied to it as a pre-processing stage. Finally, section 5 presents some conclusions of both single- and multi-channel speech enhancers in a complete speaker identification system.

2. SINGLE-CHANNEL SPEECH ENHANCEMENT TECHNIQUES

Single-channel speech enhancement techniques apply to situations in which a unique acquisition channel is available. This may be imposed by the system used (as telephone-based applications) or by the availability of the desired signal (as pre-recorded applications). When the noise process is stationary and speech activity can be detected, spectral subtraction (SS) is a direct way to enhance the noisy speech [3].

2.1. Spectral Subtraction Process

Most of the methods proposed in order to accomplish the speech enhancement process assume that the power spectral density function of the signal contaminated with incorrelated noise is equal to the power spectral density of the signal plus the power spectral density of the noisy process: this is only true in a statistical sense. Nevertheless, supposing it as a reasonable approach for the short-time spectral power function, it leads to a simple an direct way of subtracting noise from noisy speech.

Being $\overline{|R_i(\omega)|^2}$, $|Y_i(\omega)|^2$ and $|\hat{X}_i(\omega)|^2$, respectively, the power spectral estimator of the noisy process, the power spectral function of the input signal for the *i*-th analysis frame, and the power spectral estimator of the enhanced signal for the *i*-th analysis frame, the spectral subtraction process is accomplished

$$|\hat{X}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \overline{|R_i(\omega)|^2}, & \text{if } |Y_i(\omega)|^2 - \overline{|R_i(\omega)|^2} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The phase function is adjusted directly from the noisy input signal, giving the final expression for the complete enhanced signal

$$\hat{X}_i(\omega) = \left[|Y_i(\omega)|^2 - \overline{|R_i(\omega)|^2} \right]^{1/2} \cdot \frac{Y_i(\omega)}{|Y_i(\omega)|} \quad (2)$$

As it can derived from (1), the spectral subtraction method can lead to negative values, resulting from differences among the noise estimator and the actual noise frame. To cope with this problem, negative values must be set to cero, producing spectral spikes, well-known as "musical noise". This effect causes an annoying perception of enhanced speech and, therefore, it must be corrected.

2.2. Spectral Subtraction With Oversubtraction Model

This alternative to the classical spectral subtraction (SS) procedure was first introduce in order to compensate for the "musical noise" effect [4]. The general expression of the SS with oversubtraction model is given by:

$$|\hat{X}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha \cdot \overline{|R_i(\omega)|}^2, \text{if } |Y_i(\omega)|^2 - \overline{|R_i(\omega)|}^2 > \beta \cdot \overline{|R_i(\omega)|}^2 \\ \beta \cdot \overline{|R_i(\omega)|}^2, \quad \text{otherwise} \end{cases} \quad (3)$$

where $\alpha > 1$ minimizes the appearance of negative values that generate spectral spikes, and $0 < \beta \ll 1$ sets an spectral flooring which reduces the perception of musical noise. The optimal value for α can be set as a function of the SNR, as high SNR frames need less compensation that low SNR frames.

2.3. Non-Linear Spectral Subtraction

Non-Linear Spectral Subtraction (NSS) approach [5] is based in combining two different ideas: *i*) The use of an extended noise model, with an estimator of the noisy process and an oversubtraction model, and *ii)* Non-linear implementation of the subtraction process, taking into account that the subtraction process must depend on the SNR of the frame, in order to apply less subtraction with high SNRs and vice versa.

In the NSS technique, an estimate of both noise and speech can be derived from the following expresions,

$$\overline{|R_i(\omega)|} = \lambda_R \overline{|R_{i-1}(\omega)|} + (1 - \lambda_R)|R_i(\omega)| \quad (4)$$

and

$$\overline{|Y_i(\omega)|} = \lambda_Y \overline{|Y_{i-1}(\omega)|} + (1 - \lambda_Y)|Y_i(\omega)| \quad (5)$$

For the extended model of noise, it will be necessary to use a generic function $\Phi\left[\rho_i(\omega), \alpha_i(\omega), \overline{|R_i(\omega)|}\right]$ which depends on the noise estimator, on the spectral-dependent oversubtraction factor, $\alpha_i(\omega)$, and on the SNR of each spectral component of the analysis frame, $\rho_i(\omega)$, that can be calculated as

$$\rho_i(\omega) = \frac{\overline{|Y_{SNR,i}(\omega)|}}{\overline{|R_i(\omega)|}} \quad (6)$$

being

$$\overline{|Y_{SNR,i}(\omega)|} = \lambda_{SNR} \overline{|Y_{i-1}(\omega)|} + (1 - \lambda_{SNR})|Y_i(\omega)| \quad (7)$$

The function Φ is an arbitrary non linear function that encloses the subtraction process, taking into account the SNR of each spectral component, with upper and lower boundaries:

$$\overline{|R_i(\omega)|} \le \Phi\left[\rho_i(\omega), \alpha_i(\omega), \overline{|R_i(\omega)|}\right] \le 3 \cdot \overline{|R_i(\omega)|} \quad (8)$$

3. MULTI-CHANNEL SPEECH ENHANCEMENT TECHNIQUES

Multi-channel speech enhancement techniques take advantage of the availability of multiple signal input to our system, making possible the use of noise references in an adaptive noise cancellation device, the use of phase alignment to reject undesired noise components, or even the use of phase alignment and noise cancellation stages into a combined scheme [8]. We are presenting two different systems, the first of them based in adaptive noise cancellation, an the second based in speech beamforming through array processing.

3.1. Adaptive Noise Cancellation

Adaptive noise cancellation is a powerful speech enhancement technique [6] based in the availability of an auxiliar channel, known as reference path, where a correlated sample or reference of the contaminating noise is present. This reference input will be filtered following an adaptive algorithm, in order to subtract the output of this filtering process from the main path, where noisy speech is present.

The LMS algorithm is a practical algorithm that permits us to find an aproximated solution to the optimal filtering process. It has the following formulation:

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + 2 \cdot \mu \cdot e(n) \cdot \boldsymbol{y}_n \quad (9)$$

being $\boldsymbol{w}$ the vector of coefficients of the filter, $\boldsymbol{y}$ the vector reference signal, $e(n)$ the error signal and μ the adaptation constant that controls the stability and the speed of convergence of the adaptive procedure.

The process of adaptive filtering is optimal in the sense that error signal $e(n)$ guides the convergence of the whole process. Nevertheless, in practical implementations, it is very difficult to find a speech-free noise reference, and to obtain sufficient degree of correlation between reference and contaminating noises.

3.2. Multisensor beamforming

Multisensor beamforming through microphone arrays [7], derived from radar and sonar applications, can be implemented in a variety of ways, being delay-and-sum beamforming the most direct approach. The underlaying idea of this scheme is based on the assumption that the contribution of the reflexions is small, and that we know the direction of arrival of the desired signal. Then, through a correct alignment of the phase function in each sensor, the desired signal can be enhanced, rejecting all the noisy components not aligned in phase. So, for the m-th channel of the system we will have:

$$y_m(n) = x(n - \tau_m) + r_m(n) \qquad (10)$$

where $x(n)$ will be the desired signal, τ_m the delay applied to the input signal, $r_m(n)$ the noise present in the channel and $y_m(n)$ the available input of this channel. The overall output of the multisensor system will be obtained by adding all contributions, with adecuate compensating delays in each of them, giving:

$$\hat{x}(n) = {}^{1}\!/_{M} \cdot \sum_{m=1}^{M} y_m(n + \hat{\tau}_m) \qquad (11)$$

This delay and sum beamforming process is a very robust scheme. The delay estimation errors reduce the enhancement process in terms of SNR, but inducing little distortion. Anyway, there is a theoretical limit to the enhancement process: supposing we have optimal estimators for the delay in each channel, there is no room reverberation and the contributions of each channel are independent from each other, the maximum enhancement possible will be $10 \cdot \log_{10} M$ (dB), being M the number of microphones used.

4. RECOGNITION RESULTS

4.1. Speaker Identification System

Each one of the pre-processing enhancing techniques proposed have been comparatively used in a speaker identification system. This system [9] is based in ergodic HMMs, 8 states and 8 mixtures per state, trained with 60 sec. of read clean speech (SNR>30 dB) for each of the 25 male speakers involved. Speech has been acquired at 8 kHz. with 8 bits, bandlimiting it at 300-3400 kHz. (telephone-like quality). Noise has been artificially added to clean speech; two kinds of noise has been used for testing: white gaussian noise, and real fan noise extracted from a computing system, each of them added at 20, 15, 10 and 5 dB SNR. The parametric vector used is formed by 10 LPCC coefficientes, discarding c_0.

4.2. Acoustical Mismatch Among Phases

As stated previously in 4.1, the whole database has been degraded with two different types of noises (white gaussian noise and fan noise) at different SNRs (20 dB, 15 dB, 10 dB and 5 dB). The training phase has been carried out without acoustical degradation, preserving original SNR (>30 dB). Consequently, there is an acoustical mismatch between phases, and Table 1 shows the performance degradation of the speaker identification system with testing utterances of 8 sec. of duration.

ID Rate (%)	>30 dB	20 dB	15 dB	10 dB	5 dB
White gauss.	100	90.4	46.2	19.2	4.2
Fan noise	100	98.0	76.0	13.4	6.2

Table 1: Speaker ID Rate with testing utterances degraded with white gaussian or fan noises, at different SNRs.

4.2. Single-Channel Speaker Identification

In order to enhance the speech entering our recognition system, we have applied as an acoustical pre-processing stage the three spectral subtraction derived algorithms stated respectively in 2.1, 2.2 and 2.3, namely classical spectral subtraction, spectral subtraction with oversubtraction model and non-linear spectral subtraction. Table 2 shows the results obtained, and Figures 1 and 2, show graphically these results, with regard to the kind of degrading noise employed in each case.

ID Rates (%)	20 dB		15 dB		10 dB		5 dB	
	W	F	W	F	W	F	W	F
SS	96.4	97.4	84.2	94.2	31.0	69.4	9.2	24.6
SS+Over.	94.4	98.8	89.4	93.4	40.4	71.8	10.4	31.2
NSS	77.6	98.0	66.2	93.4	30.6	73.6	9.8	35.8

Table 2: Speaker ID rates, when Spectral Sub. (SS), SS with Oversub. model (SS+Over.) and Non-linear SS (NSS) are used for white and fan noises at different SNRs.

4.3. Multi-Channel Speaker Identification

Multi-channel enhancement has been carried out, using adaptive noise cancellation and delay-and-sum speech beamforming. The adaptive noise cancellation system has been artificially implemented through the simulation of the impulse responses of a room using a geometrical approach to room acoustics design. These responses had been used to filter speech coming from one point of the room and noise coming from another point of it.

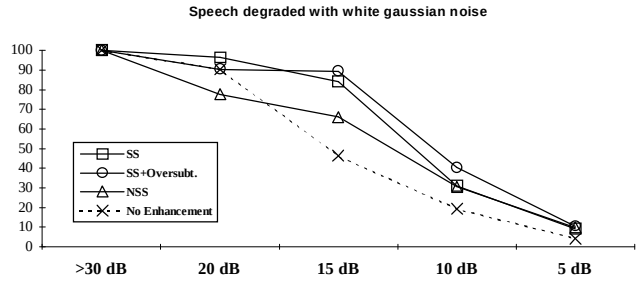

Figure 1: Results presented in Table 2 for white gaussian noise.

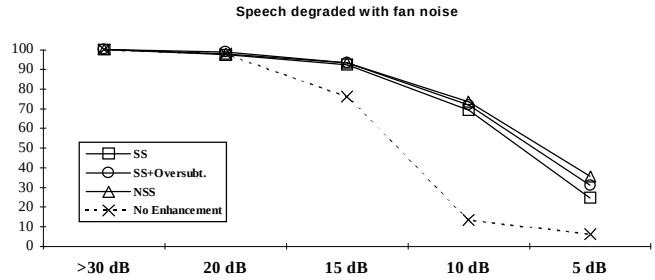

Figure 2: Results presented in Table 2 for fan noise.

Consequently, noise has been added to reverberant speech in order to obtain the required SNR, and this noisy reverberant signal has been used in the main path. In the reference path, the original noise signal has been used.

For the speech beamformer, a low-complexity four microphone array has been used, simulating the impulse responses for noise and speech entering each one of the microphones employed. This artificial procedure has permitted to obtain directly the delay corresponding to each of the four paths involved in the system.

Results on each multi-channel approach, regarding the kind of noise used, are presented in Figures 3 and 4.

5. CONCLUSIONS

Acoustical mismatch among training and testing phases degrades outstandingly speaker identification results. Enhancement techniques applied, as pre-processing stages, to speaker ID systems remarkably improve recognition results. Single-channel enhancing techniques, based on spectral subtraction as a direct procedure to implement, produce good recognition results when acoustical degradation stands over 10 dB SNR, though introducing apreciable distortion on the recovered speech.

Multi-channel speech enhancement systems produce excellent results for moderate and high noise levels (SNR>5 dB). Adaptive cancellation outperforms any other technique, with excellent results even for SNR=5dB. Anyway, this technique is not much realistic, as reference path must be signal free for real applications. Array processing is a very useful technique, and excellent results can be obtained for SNR>5dB in a very realistic manner.

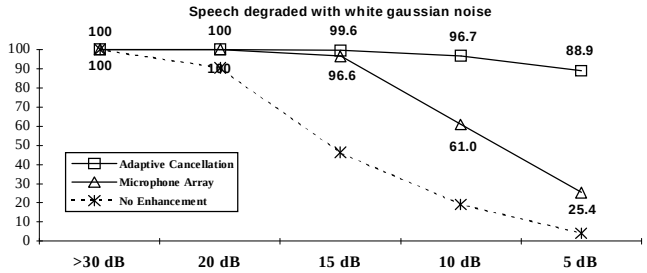

Figure 3: Multi-channel ID results for white gaussian noise.

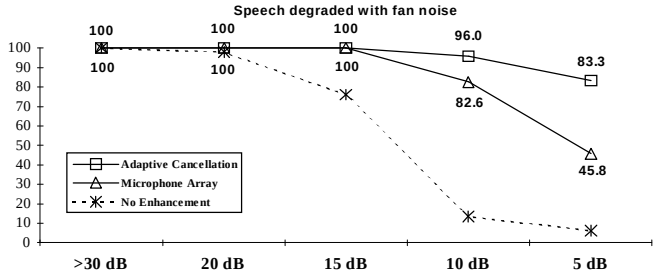

Figure 4: Multi-channel ID rates for fan noise.

REFERENCES

[1] S. Furui, "Towards Robust Speech Recognition Under Adverse Conditions", ESCA Workshop on Speech Proc. in Adverse Conditions, pp. 31-42, 1992.

[2] J. Ortega-García et al., "Robust Speech Modeling for Speaker ID in Forensic Acoustics", ESCA Workshop on Automatic Speaker Recognition, pp. 217-220, 1994.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. ASSP* 27(2): 113-120, April 1979.

[4] M. Berouti et al., "Enhancement of Speech Corrupted by Acoustic Noise", Proc. ICASSP, pp. 208-211, 1979.

[5] P.Lockwood et al., "Experiments with a Nonlinear Spectral Subtractor ...", *Speech Communication* 11: 215-228, 1992.

[6] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.

[7] Q. Lin, E. Jan and J. Flanagan, "Microphone Arrays and Speaker Identification", *IEEE Trans. SAP* 2(4): 622-629, October 1994.

[8] J. González-Rodríguez et. al., "Increasing Robustness in GMM Speaker Recognition Systems with Low Complexity Microphone Arrays", Proc. ICSLP, somewhere in these Proceedings, 1996.

[9] J. Ortega-García, *Speech Enhancement Techniques Applied to Speaker Recognition Systems* (in Spanish), Ph. D. Thesis, Univ. Politécnica de Madrid, Spain, 1996.