PROVIDING SINGLE AND MULTI-CHANNEL ACOUSTICAL ROBUSTNESS TO SPEAKER IDENTIFICATION SYSTEMS

Javier Ortega-Garcia and Joaquin Gonzalez-Rodriguez

Dept. de Ingeniería Audiovisual y Comunicaciones
E.U.I.T. Telecomunicación - Universidad Politécnica de Madrid
Crta. Valencia km. 7, Campus Sur, E-28031 Madrid, Spain
jortega@diac.upm.es

ABSTRACT

Acoustical mismatch between training and testing phases induce degradation of performance in automatic speaker recognition systems [1,2]. Providing robustness to speaker recognizers has to be, therefore, a priority matter. Robustness in the acoustical stage can be accomplished through speech enhancement techniques as a prior stage to the recognizer. These techniques are oriented to the reduction of the impact that acoustical noise produces on the input signal [3,4]. In this paper, several spectral subtraction-derived techniques are used to enhance single-channel noisy speech. Other perspectives, based in dual-channel (adaptive filtering) and multi-channel (microphone arrays) processing are also presented as optimal solutions to speech enhancement needs. A comparative analysis of the proposed 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

The identification of the talker is a growing necessity in many fields of application of speech technologies, specially within the framework of security (remote) applications. A mismatch between training and testing conditions induces a severe degradation in the performance of those systems. This question has restrained the development of real-world non-specific applications, as testing conditions may be unknown during the training process, done under ideal laboratory conditions.

Providing robustness, that is, reducing the degradation of performance due to the mismatch between phases, can be accomplished, in a general manner, 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 exhibit immunity to the noise source and *iii*) the modeling stage, combining adequate models of noise and clean signal in order to recognize noisy speech.

In this paper, a wide analysis of single (section 2) and multi-channel (section 3) techniques providing robustness in the acoustical stage to ergodic HMM-based speaker identification systems (section 4) is presented, leading to some relevant conclusions (section 5).

2. SINGLE-CHANNEL SPEECH ENHANCEMENT TECHNIQUES

Many applications (telephone-based, pre-recorded samples, etc.) apply to situations in which a unique acquisition channel is available. When the noise is stationary and speech activity can be detected, spectral subtraction (SS) is a direct way to enhance noisy speech [5], giving rise to what from here on are called single-channel enhancement techniques.

2.1. "Classical" Spectral Subtraction Technique

In the power spectral density domain, we may assume, in order to accomplish the speech enhancement process, 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: however, this is only true in a statistical sense. Anyway, for the short-time spectral power function, we may suppose it as a reasonable approach, leading 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 through the equation:

This work has been supported by the CICYT under Project TIC 94-0030

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

where the phase function is adjusted directly in the enhanced signal from the noisy input signal. As it can derived from the upper part of (1), the spectral subtraction method can lead to negative values, resulting from negative differences between the noise estimator and the actual noise value. To cope with this problem, negative values must be set to zero, giving rise to the well-known "musical noise" effect, consisting in sudden spectral spikes. This kind of noise causes an annoying perception of the enhanced speech and, therefore, it must be corrected.

2.2. Spectral Subtraction with Oversubtraction Model

In order to correct the appearance of the "musical noise", it was proposed [6] to use an oversubtraction model for the noise, given by:

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

where $\alpha > 1$ minimizes the appearance of negative values that generate spectral spikes, and $0 < \beta << 1$ 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 [7] is based in the combination of two different ideas: *i*) The use of an extended noise model, with an estimator of the noisy process and an oversubtraction model, and *ii)* a non-linear implementation of the subtraction process, taking into account that the subtraction process must depend on the SNR of the spectral components of each analysis frame, in order to apply less subtraction with high SNRs and vice versa.

A generic function $\Phi\left[\rho_i(\omega), \alpha_i(\omega), \overline{\left|R_i(\omega)\right|}\right]$ will be necessary for the extended model of noise; this function will depend 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:

$$\rho_i(\omega) = \frac{\overline{\left|Y_{SNR,i}(\omega)\right|}}{\overline{\left|R_i(\omega)\right|}} \quad (3)$$

being:

$$\overline{\left|Y_{SNR,i}(\omega)\right|} = \lambda_{SNR} \overline{\left|Y_{i-1}(\omega)\right|} + (1 - \lambda_{SNR}) \left|Y_i(\omega)\right| \quad (4)$$

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:

$$\overline{\left|R_i(\omega)\right|} \le \Phi\left[\rho_i(\omega), \alpha_i(\omega), \overline{\left|R_i(\omega)\right|}\right] \le 3 \cdot \overline{\left|R_i(\omega)\right|} \quad (5)$$

2.4. Sub-band Non-Linear Spectral Subtraction

Sub-band NSS (SB-NSS) consists basically in the application of the non-linear spectral subtractor in 1/3 octave sub-bands, instead of using an oversubtraction factor for each spectral component [4]. The equivalent level in the actual analysis band is derived from

$$B_L = 10 \cdot \log_{10}\left(\sum_{i=1}^{N} 10^{L_{pi}/10} \cdot \Delta f_i\right) \quad (6)$$

where Δf_i is the spectral resolution used, and N the number of components in the actual analysis band. 21 standard 1/3 octave bands have been used, from 31.25 Hz to 4 kHz.

This procedure reduces considerably the computational load required for the NSS procedure as the spectral components needed to accomplish the NSS process are now reduced to 21 values.

3. MULTI-CHANNEL SPEECH ENHANCEMENT TECHNIQUES

If we are able to have several input channels to our application and we may control the arrangement of them, we can take advantage of the availability of multiple signal inputs to our system using multi-cannel speech enhancement techniques, being the most direct of these: *i*) the use of noise references in an adaptive noise cancellation device, *ii)* the use of phase alignment to reject undesired noise components, or even *iii)* the use of phase alignment and noise cancellation stages into a combined scheme [8].

3.1. Adaptive Noise Cancellation

Adaptive noise cancellation is a powerful speech enhancement technique [9] based in the availability of an auxiliary 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 implementation of this adaptive process that permits us to find an approximated solution to the optimal filtering process. It has the following formulation:

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + 2 \cdot \mu \cdot e(n) \cdot \boldsymbol{y}_n \quad (7)$$

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 [10], derived from radar and sonar applications, can be implemented in a variety of ways, being delay-and-sum beamforming the most direct approach. The underlying idea of this scheme is based on the assumption that the contribution of the reflections 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:

$$y_m(n) = x(n - \tau_m) + r_m(n) \quad (8)$$

where $x(n)$ will be the desired signal, τ_m the delay applied to the input signal for a correct phase alignment, $r_m(n)$ 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 adequate compensating delays in each of them, giving:

$$\hat{x}(n) = 1/M \cdot \sum_{m=1}^{M} y_m(n + \hat{\tau}_m) \quad (9)$$

This delay and sum beamforming process is a very robust scheme. The delay estimation errors reduce the enhancement process in terms of SNR, although inducing only a low degree of distortion.

4. SPEAKER IDENTIFICATION RESULTS

4.1. Overall System Description

The baseline speaker identification system used [4] 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., encoded with 8 bits and μ-law, and bandlimited in the range 200-3400 kHz. in order to obtain telephone-like quality.

The training phase has been carried out without acoustical degradation, preserving the original SNR (>30 dB). For the testing phase, noise has been artificially added to clean speech; three kinds of noise has been used for testing: white gaussian noise, real fan noise extracted from a computing system, and tonal noise, consisting of tones at 250, 500, 1k, and 2k Hz.; these noises have been added to speech at 20, 15, 10 and 5 dB SNR. Each one of the pre-processing enhancing techniques proposed in sections 2 and 3 have been comparatively used. In all cases, the parametric vector used is formed by 10 LPCC coefficients, discarding c_0.

4.2. Single-Channel Speaker Identification

The single-channel enhancement techniques described in section 2 (namely, "Classical" Spectral Subtraction, Spectral Subtraction with Oversubtraction model, Non-Linear Spectral Subtraction and Sub-Band Non-Linear Spectral Subtraction) are applied as an acoustical pre-processing stage to the speaker recognizer described in 4.1. Table 1 shows the results obtained.

Single-Channel Enhancement Techniques						
SNR	*Noise*	*NE*	*S1*	*S2*	*S3*	*S4*
20 dB	*W*	90.4	96.4	94.4	77.6	58.8
	F	98.0	97.4	98.8	98.0	93.0
	T	95.2	96.0	100	99.6	97.6
15 dB	*W*	46.2	84.2	89.4	66.2	45.6
	F	76.0	92.4	93.4	93.4	84.0
	T	48.4	100	97.2	97.4	89.4
10 dB	*W*	19.2	31.0	40.4	30.6	21.6
	F	13.4	69.4	71.8	73.6	64.6
	T	15.6	84.0	70.2	81.2	67.8
5 dB	*W*	4.2	9.2	10.4	9.8	9.4
	F	6.2	24.6	31.2	35.8	25.0
	T	7.4	98.2	27.0	30.4	27.6

Table 1: *Speaker ID rates (%), when No Enhancement (NE), "Classical" Spectral Subtraction (S1), Spectral Subtraction with Oversubtraction Model (S2), Non-linear Spectral Subtraction (S3), and Sub-Band Non-Linear Spectral Subtraction (S4) are used, with white (W), fan (F) and tonal (T) noises at the stated SNRs.*

4.3. Multi-Channel Speaker Identification

The two multi-channel speech enhancement techniques proposed in section 3 (namely, the adaptive noise canceller and the delay-and-sum beamformer) have been implemented in order to obtain comparative results with regard to single-channel enhancement techniques.

The adaptive noise cancellation system (described in 3.1.) has been simulated through the impulse response 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. 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 (described in 3.2.), a low-complexity four microphone linear 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 Table 2.

Multi-Channel Enhancement Techniques				
SNR	*Noise*	*NE*	*ANC*	*MA*
20 dB	*W*	90.4	100	100
	F	98.0	100	100
	T	95.2	100	100
15 dB	*W*	46.2	99.6	96.6
	F	76.0	100	100
	T	48.4	100	98.0
10 dB	*W*	19.2	96.7	61.0
	F	13.4	96.0	82.6
	T	15.6	100	84.0
5 dB	*W*	4.2	88.9	25.4
	F	6.2	83.3	45.8
	T	7.4	100	54.4

Table 2: *Speaker ID rates (%), when No Enhancement (NE), Adaptive Noise Cancelling (ANC) and low-complexity Microphone Array Processing (MA) are applied, with white (W), fan (F) and tonal (T) noises at the stated SNRs.*

5. CONCLUSIONS

The baseline results (*NE*) in Tables 1 and 2 demonstrate that acoustical mismatch among training and testing phases degrades outstandingly speaker identification results. On the other hand, speaker ID results remarkably improve when enhancement techniques are applied as pre-processing stages. Single-channel enhancing techniques produce good recognition results when acoustical degradation stands over 10dB SNR, specially for real fan and tonal noises, though introducing some level of 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 completely realistic, as reference path must be signal free for real applications, and the correlation between paths restricts noise cancellation process.

On the contrary, low-complexity array processing is a very promising technique, as excellent results can be obtained for SNR>5dB with not much implementation constraints in a very realistic manner. For practical systems, the time-delay estimation of each path is the main problem to be solved, knowing that the estimation error will only affect on the SNR obtained, without inducing additional distortion.

REFERENCES

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

[2] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition -Fundamentals and Applications*, Kluwer Academic Publishers., Dordrecht, NL, 1996.

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

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

[5] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on ASSP*, Vol. ASSP-27, No. 2, pp. 113-120, April 1979.

[6] M. Berouti *et al.*, "Enhancement of Speech Corrupted by Acoustic Noise", Proc. ICASSP-79, pp. 208-211, 1979.

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

[8] J. Gonzalez-Rodriguez *et. al.*, "Increasing Robustness in GMM Speaker Recognition Systems with Low Complexity Microphone Arrays", Proc. Intl. Conf. Spoken Language Proc., ICSLP-96, pp. 1333-1336, Philadelphia.

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

[10] Q. Lin, E. Jan and J. Flanagan, "Microphone Arrays and Speaker Identification", *IEEE Trans. on SAP*, Vol. SAP-2, No. 4, pp. 622-629, October 1994.