INCREASING ROBUSTNESS IN GMM SPEAKER RECOGNITION SYSTEMS FOR NOISY AND REVERBERANT SPEECH WITH LOW COMPLEXITY MICROPHONE ARRAYS

Joaquín González-Rodríguez[(1)], Javier Ortega-García[(1)], César Martín[(2)] and Luis Hernández[(2)]

(1) DIAC - EUIT Telecomunicación, Universidad Politécnica de Madrid
(2) GAPS - SSR - ETSIT, UPM
Ctra. Valencia, Km.7, Campus Sur, E-28031, Madrid, España
email: jgonzalz@diac.upm.es

ABSTRACT [(*)]

In this paper we describe the additive robustness obtained through the combined use of a first acoustic processing step based on a low complexity microphone array, followed by a spectral normalization step. Microphone arrays have shown to provide good results in reducing different sources of acoustic degradation. However, microphone arrays produce linear filtering effects that need to be compensated in order to obtain a minimal spectral distortion. In this contribution we will present the combination of a microphone array together with different well known spectral normalization techniques as preprocessing stages to a Gaussian Mixture Models (GMM) based text-independent speaker recognition system. We will show that the combination of these extensively used techniques in the fields of speech enhancement and robust speaker recognition respectively, greatly improves the results obtained when the system is tested in noisy reverberant environments with short utterances from unconstrained conversational speech.

1. INTRODUCTION

There are a lot of real world situations where the use of a close talking microphone is not desirable or even possible, such as an access control system, where the position of the claiming speaker is neither known nor fixed during the utterance, or command and control systems where an operator needs to move around a room without restrictions. In such cases, in addition, there are usually competing noises that will disturb the signal picked up by the microphone, together with the reverberation introduced due to the multipath propagation from the sources (speech and noises) to the receiver [1], as we can see in figure 1. Though this problem has been addressed previously, the system proposed in [2] consisted in large complexity microphones arrays (tens of microphones in any case) as an acoustic preprocessor to a VQ based speaker recognition system. The overall system proposed in this contribution looks for robustness at low computational and hardware joint cost. Instead of complex projection techniques in the cepstral space [3] or important modifications in the HMM architectures [4], we place an speech enhancement stage dealing with noisy reverberant speech [5] and a simple spectral normalization technique [6], to compensate for the effects introduced in the acoustic stage. Of course those effects will not be completely cancelled, but this rather simple system will work far better than the mono channel original system.

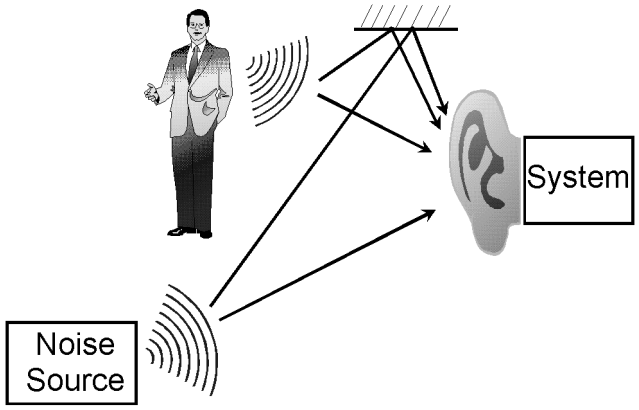

Figure 1: Speech and noise sources at the receiver with their respective transfer functions.

The speaker recognition system works with one Gaussian Mixture Model (1-state HMM) per speaker [7], which have been shown to be robust and an excellent model of the speaker short-time characteristics, outperforming VQ and ergodic HMM systems [8].

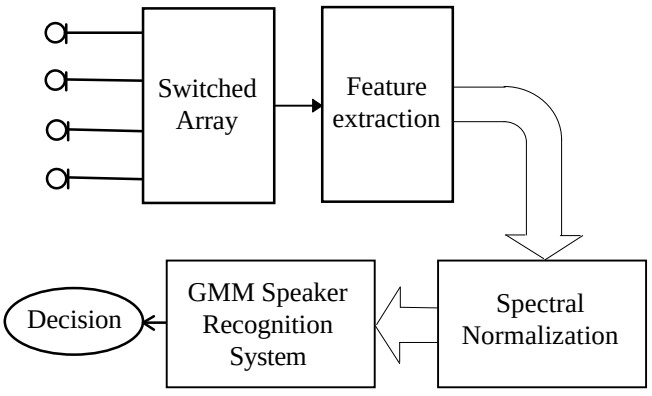

Figure 2: Overview of the whole system, where the array structure is described in figure 3.

[(*)] This work has been funded by CICYT under Project TIC94-0030

2. SYSTEM DESCRIPTION

The array structure used is that described in [5], where the adaptation of a two stage system (see figure 3) is switched with a speech/pause detector. The first stage is that involved with the beamforming of the array, and is readapted when speech is present. When no speech is detected, the delay estimates are not changed and the adaptive filters coefficients are readapted, acting as adaptive noise cancellers.

The time delay estimates applied to each channel are estimated through temporal crosscorrelation between the corresponding channels, with a plausability check of that result to avoid obviously incorrect estimations, such as rapid movements of the speaker. The adaptive filters work under a conventional LMS algorithm, while the speech/pause detector is a very simple one based on two underestimated thresholds over the short time energy, where no adaptation is accomplished when neither speech nor pause is clearly detected.

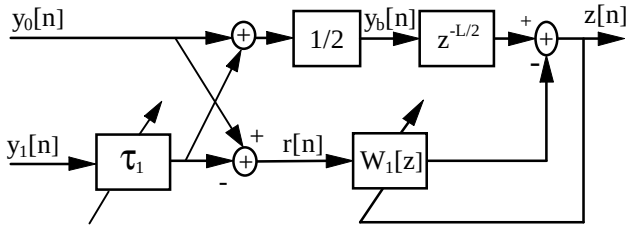

Figure 3: Two-stage 2-channel switched array.

The beamformed and final output are given by the equations:

$$y_b[n] = 0.5 \bullet \{ y_0[n] + y_1[n - \tau_1] \}$$

$$z[n] = y_b[n - L/2] - \{ y_0[n] - y_1[n - \tau_1] \} * w_1[n]$$

The acoustic characteristics and capabilities of the array are determined by the election of the number of microphones involved, and their position. According to the intention of having a rather simple array, we decided to work with a broadband linear array of 4 microphones equally spaced 10 centimeters one from each other (our maximum frequency considered is 4 KHz). This of course has it limitations, as can be clearly seen in figure 5, but the linear effects introduced are expected to be minimized at the spectral normalization stage of the recognizer.

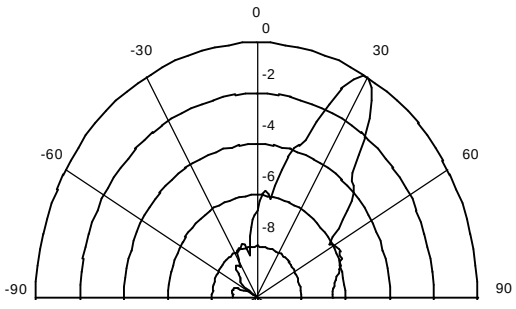

Figure 4: Average directivity pattern for the octave frequencies 250, 500, 1000 and 2000 Hz for a 4 microphone linear array equally spaced 10 centimeters one from each other, for an estimated angle of arrival of 30º.

As we can see in the following figure, a secondary lobe appears for frequencies over 2000 Hz, but this of course is a compromise between the simplicity of the system (4 microphones, broadband, total lenth of 40 centimeter) and the ideal acoustical capabilities of the system.

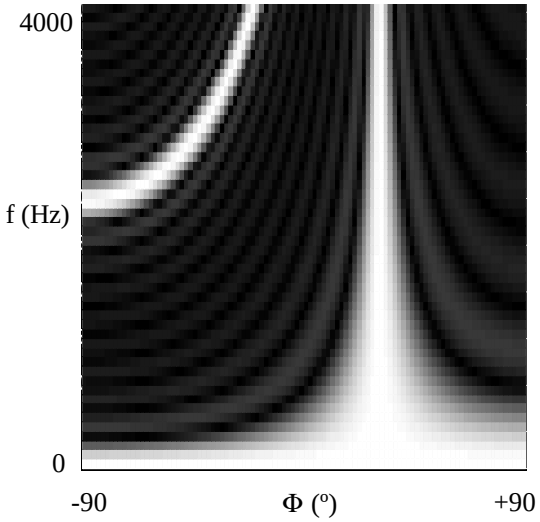

Figure 5: 2-D view of the directivity patterns for each frequency below 4 Khz, for an estimated angle of arrival of 30º.

The spectral normalization techniques used at the output of the array structure are the well known Cepstral Mean Normalization (CMN) and RASTA processing [6], which have been shown effective compensating for the linear effects introduced in the channel.

In CMN the mean of the cepstral vectors is subtracted in order to high-pass filter the original cepstral coefficients:

$$y[n] = x[n] - \frac{1}{N}\sum_{i=1}^{N} x_i[n]$$

RASTA processing of speech again high-pass filter the cepstral coefficients with the following difference equation:

$$y[n] = x[n] - x[n - 1] + 0.97\, y[n - 1]$$

The speaker recognition system models the speaker characteristics with a one state model per speaker with a discrete set of gaussian mixtures (M=8, M=16 or M=32) corresponding to the probabilistic distribution of the LPCCepstrum vectors obtained from the speaker database described below.

In GMM systems, each speaker model λ is given by:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \qquad i = 1, \ldots, M$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i ; a gaussian mixture density is given by a weighted sum of component densities:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

where $\vec{x}$ is our L-dimensional cepstral vector, with mixture weights p_i and component densities $b_i(\vec{x})$ given by the equation:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{L/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\ \Sigma_i^{-1}\ (\vec{x}-\vec{\mu}_i)\right\}$$

3. EXPERIMENTS

3.1. Speaker Database

The speech data have been extracted from the DIAC2 speaker database, recorded at DIAC-EUIT Telecom. UPM, consisting in several minutes of unconstrained speech from each one of 25 male speakers, recorded with a high quality close talking microphone in a quiet studio (SNR>30 dB). The database has been labelled as speech/silence by direct observation and listening of the files.

3.2. Multipath propagation in reverberant rooms

The impulse response between any two points of a room is simulated in a computer through the image method, according to the acoustic ray theory [9], choosing the form, dimensions, absorption coefficients of the walls, and maximum reflexion order. In this experiment, we choose a room of 6x4x3 meters, placing the speech source at about 1.5 meters of the array (and about 30º off-axis), and the noise source (white noise) at the other side of the room, at about 4.5 meters (about -15º off-axis). We then calculate the 8 impulse responses from each of the two sources to the four microphones of the line array, equally spaced 10 cm, as shown by the multipath propagation of the different rays in fig. 6.

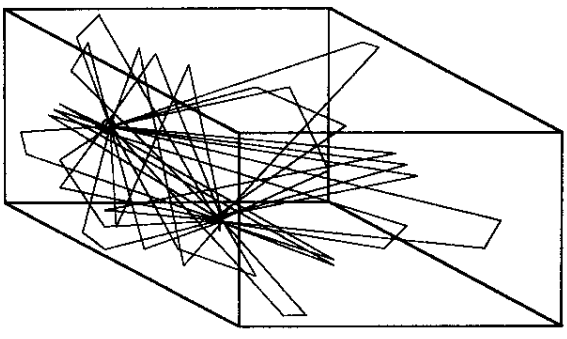

Figure 6: Multipath propagation through the images method between any two points of a room.

The impulse responses obtained have the form of that in figure 7 where the ray attenuations are calculated with respect to the direct path arrival echo. The reflexion order can be truncated at any integer, obtaining similar results for reflexions above the third order.

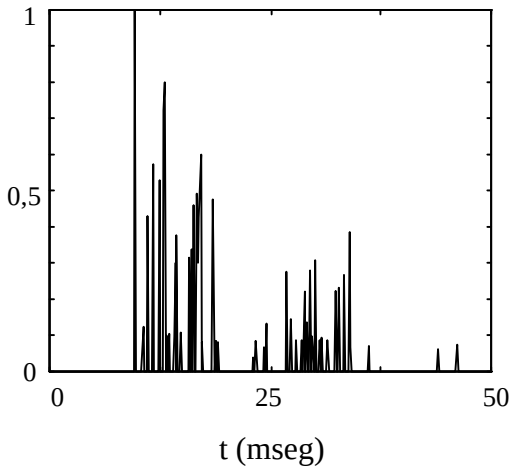

Figure 7: Impulse response obtained with the room simulation program. The echo amplitudes of the impulse responses obtained are relative to that of the direct path.

3.3. Training and testing

We train our 25 male speaker models in clean conditions (without noise or reverberation) with 14 seconds of actual speech (silences removed) per speaker from the speech database, with various values of M, the number of gaussian densities. This process is repeated when the CMN or RASTA models are to be used, obtaining three types of 'clean' models (no processing, CMN and RASTA).

In the testing stage, we artificially generate the input signal to each of the microphones adding two convolutioned signals (one for the speech and another for the noise with their respective impulse responses) at two different SNR, measured as the ratio between the average energy at speech frames to that at noisy frames, where the noise source is white noise. We test the system with 30 miliseconds LPCCepstrum vectors from 10 overlapping segments of 5 seconds in 4 different situations corresponding to the different stages in the processing (clean speech, input to one microphone of the array, beamformed signal, and output signal from the array processing system). The recognized speaker is that of the highest output probability without any type of postprocessing.

4. RESULTS

When the system recognizes the clean speech segments (SNR>30dB) with the clean models (original, CMN or RASTA), and no room simulation is performed, it obtains recognition rates above 96 % for any value of M (8,16 or 32).

The system capabilities have been tested at two different input SNR (5 and 10 dB). The following tables and curves summarize the results obtained in both cases and for different values of M, the number of gaussian mixtures in each speaker model.

SNRin = 5 dB			
M (gauss. mix.) = 8			
Normalization:	None	CMN	RASTA
Microph. #1	9.6	32.4	20.4
Beamformed	23.6	46.8	49.2
Output	60.8	53.2	52.8

Table 1: Recognitions results with the whole system, with a low input SNR.

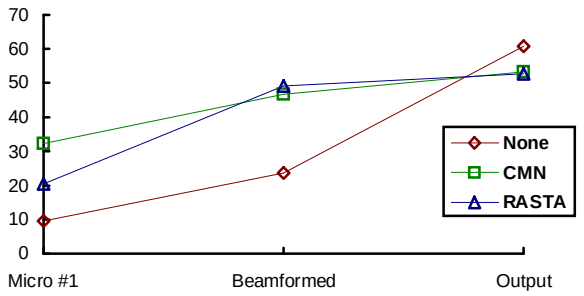

Figure 8: Graphical representation of speaker recognition results of table 1 .

SNRin = 15 dB			
M (gauss. mix.) = 8			
Normalization:	None	CMN	RASTA
Microph. #1	64.0	80.0	68.4
Beamformed	92.0	97.0	86.4
Output	96.0	97.0	86.4
M (gauss. mix.) = 16			
Normalization:	None	CMN	RASTA
Microph. #1	65.2	78.4	80.0
Beamformed	100.0	99.6	99.2
Output	100.0	99.6	99.2
M (gauss. mix.) = 32			
Normalization:	None	CMN	RASTA
Microph. #1	54.8	71.2	72.8
Beamformed	99.6	96.8	94.8
Output	99.6	96.8	94.8

Table 2: Recognition results with the whole system working for a medium input SNR.

5. CONCLUSION

In this contribution we have shown how we have increased the robustness to additive and convolutional noise of a Gaussian Mixture Model speaker recognition system through the joint use of a simple four microphones switched array and a spectral normalization step to compensate the linear filtering effects introduced in the acoustic processing step. Though these are preliminary results, from the experiments described above we see the enormous potential of microphone arrays as acoustic preprocessing stages in combination of spectral normalization techniques providing robustness to speech or speaker recognitions systems.

ACKNOWLEDGEMENTS

We thank Manuel Sobreira, now at Universidad de Vigo (Spain), for the rooms acoustic analysis software.

6. REFERENCES

[1] J. Flanagan, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", J. Acoust. Soc. Am., Vol. 78, pp. 1508-1518, 1985.

[2] Q. Lin, E. Jan, and J. Flanagan, "Microphone Arrays and Speaker Identification", IEEE Trans. on Speech and Audio Processing, October 1994.

[3] Acero, A. and Stern, R.M., "Robust Speech Recognition by Normalization of the Acoustic Space", Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Proc.., pp. 893-896, 1991.

[4] Gales, M.J.F. and Young, S.J., "Cepstral Parameter Compensation for HMM Recognition in Noise", Speech Communication, n.12, pp. 231-239, 1993.

[5] D. Van Compernolle, "Speech Recognition in Noisy Environments with the Aid of Microphone Arrays", Speech Comm. 9, pp. 433-442, 1990.

[6] Junqua, J.C. and Haton, J.P., "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, Chapter 8, pp. 233-272, 1996.

[7] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. on Speech and Audio Processing, January 1995.

[8] J.Ortega-Garcia and J.Gonzalez-Rodriguez, "Comparative Performance of Automatic Speaker Identification Systems", ACUSTICA - acta acustica, Vol. 82 (1996) Suppl. 1, pp. S231.

[9] J.B. Allen and D.A. Berkley, “Image method for efficiently simulating small-room acoustics”, J. Acoust. Soc. Amer., Vol. 65, No. 4, pp. 943-950, 1979.