ROBUST SPEAKER RECOGNITION THROUGH ACOUSTIC ARRAY PROCESSING AND SPECTRAL NORMALIZATION

Joaquin Gonzalez-Rodriguez and Javier Ortega-Garcia

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
jgonzalz@diac.upm.es

ABSTRACT*

The development of a robust speaker recognition system obtained through the joint use of acoustic array processing and spectral normalization as input to a Gaussian Mixture Model speaker recognition system is described in this paper. Results obtained with these techniques have been reported previously by the authors [10], but operational problems appear if extensive testing with different configurations and testing conditions are intended. In this paper, we describe an open system that has been developed to cope with this problem. The number and geometry of the microphones, the time delay estimation method, the array processing structure and the spectral normalization technique together with the room size, noise type and SNR are some of the options that can be easily changed. It will also allow testing with real multichannel databases and any new algorithm can easily be incorporated to the system.

1. INTRODUCTION

The use of microphone arrays as input stage to speech/speaker recognition systems have been shown very effective in reducing the effects introduced by noise and reverberation [1]. However, the array introduces a filtering effect that is needed to compensate. The baseline system we have designed used the switched array structure described in [2] as input to the parametrization stage, where a spectral normalization technique was applied. There are several succesful techniques described in the literature to compensate for the channel effects [3], from which we have selected two of them, namely CMN and RASTA, to test our system under noisy and reverberant conditions. The speaker recognitions system works with Gaussian Mixture Models [4], who have been shown a robust model of the speaker characteristics [5].

Several encouraging results have been reported by the authors [10], showing the effectiveness of this approach. However, our system was developed 'cutting and pasting' previous smaller systems as the room simulation and the acoustic array processing softwares, running even in different platforms. As we can easily observe, this situation made difficult to reconfigure the system for any different structure or algorithm. In order to avoid this problem and to allow extensive testing of different geometrical structures, several array processing and spectral normalization techniques, or even other speaker or speech recognition algorithms, a new system has been developed, allowing us to easily modify any parameter of the system, evaluate new algorithms or testing in other acoustical environments. In this way, the room simulation software has been completely reprogrammed following [6], improved time delay estimation has been incorporated through the (optional) use of interpolation/decimation and coherence based delay estimation [7], and other array processing techniques as [8] are being added to the system.

2. BASELINE SYSTEM

2.1. Description

A full description of the baseline system we have used can be found in [10], which we can see in figure 1. The array structure is that described in [2], where the adaptation of a two stage system 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 plausibility check of that result to avoid obviously incorrect estimations,

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

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. 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, but the linear effects introduced are expected to be minimized at the spectral normalization stage of the recognizer.

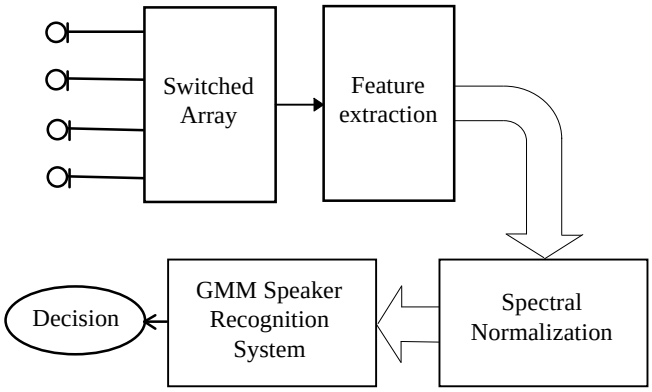

Figure 1: Graphical representation of the baseline speaker recognition system

The spectral normalization techniques used at the output of the array structure are the well known Cepstral Mean Normalization (CMN) and RASTA processing [3], which have been shown effective compensating for the linear effects introduced in the channel. 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.

2.2. Experiments

2.2.1. 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 labeled as speech/silence by direct observation and listening of the files.

2.2..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 [6], choosing the form, dimensions, absorption coefficients of the walls, and maximum reflection 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 30° off-axis), and the noise source (white noise) at the other side of the room, at about 4.5 meters (about -15° 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. These impulse reponses will be used to convolve and sum the database speech with the white noise, obtaining in each case the four channel inputs to the system.

2.2.3. 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.

2.3. Speaker Identification Results

When the system recognizes the clean speech segments from the database (SNR>30dB) with the clean models (original, CMN or RASTA), and no

room simulation is performed, recognition rates above 96% are obtained for any value of M (8,16 or 32). The system capabilities have been tested at two different input SNR (5 and 15 dB). The following table and graph shows the results obtained for input SNR=15 dB and M=8, the number of gaussian mixtures in each speaker model.

SNRin = 5 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

Table 1: Speaker recognition results for M=8 and input SNR=15 dB

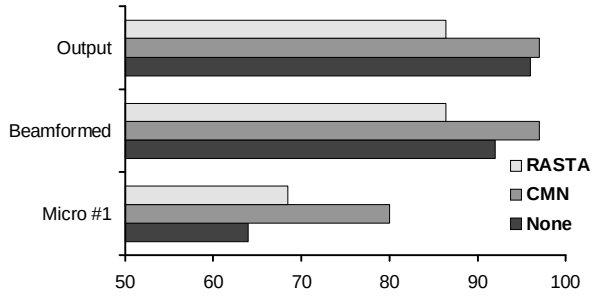

Figure 2: Graphical representation of the results shown in table 1

3. THE PROCAR SYSTEM

In this section, we will describe the objectives aimed with the design of this system (whose name comes from 'array processing' in spanish), its internal structure and we will have a brief overview of the options it has available up to date.

When using and improving the baseline system described in the previous section, we saw several drawbacks due to its closed structure so when we wanted to modify the system, we had to build a whole new system from the previous pieces. In order to cope with this problem, and observing that we were spending more time changing the system than obtaining new results, we decided to build a new open system, able to incorporate any new algorithm in any of the phases of the processing, running in a single platform, and the most important of all, making possible to run a whole experiment from the very beginning to the end with a single script, or a set of them.

The Gaussian Mixture Model speaker recognizer we are using is running over HTK [9] on a UNIX workstation, and the DIAC2 speaker database is bandpass filtered and recorded in SUNAU8 files (8 Khz, μ-law). So we decided to develop the whole system on a UNIX environment, reprogramming the room acoustics software and the array algorithms previously implemented in a PC platform.

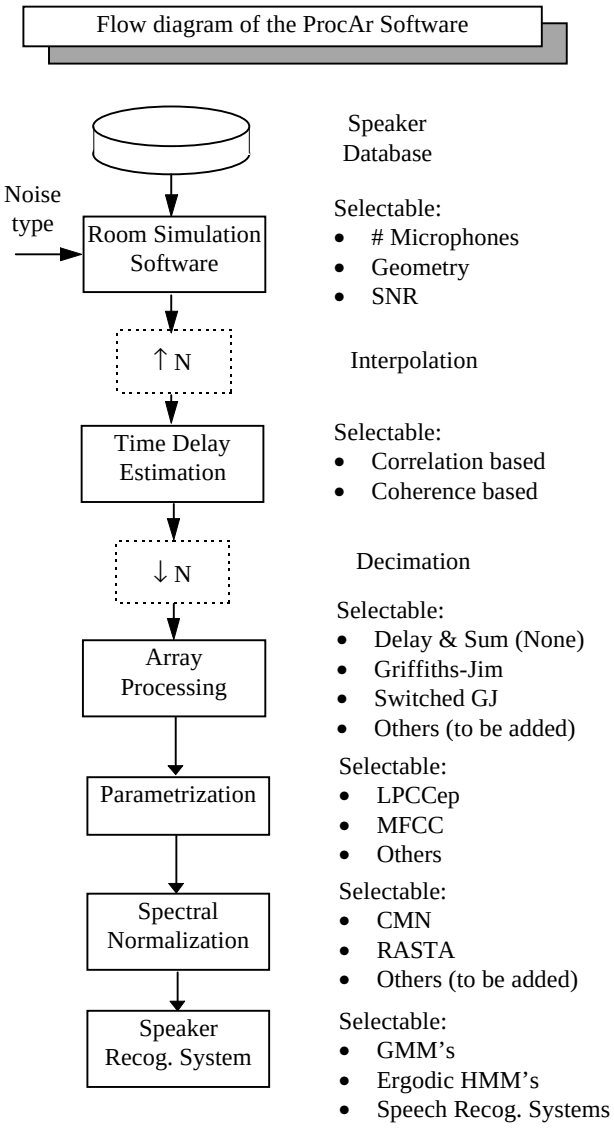

Figure 3: Modular description of the ProcAr system for extensive testing under different conditions and/or algorithms

We can see the structure of the system in figure 2. It is designed in a fully modular way, where each of the modules can be configured, suppresed or combined with other stages in a single one. The files format is the same across the whole system (HTK), which is a great advantage over the previous system. Any new test over new acoustic conditions, new databases (multichannel recorded or simulated), or

different algorithms in any of the stages can easily be run simply modifying the configuration scripts. Even new tasks, as speech recognition in noisy and reverberant conditions can be easily incorporated to the system.

Some of the routines have been completely reprogrammed, as the room acoustics software. The room transference function between any two points of a room is estimated following the algorithm described in [6]. We can see in figure 3 three examples of the same transference function between two points estimated with 512, 1024 and 2048 points respectively (64, 128 and 256 miliseconds), which can be used to study the effect over the recognition results of the transference function length on simulated data when comparing with real data.

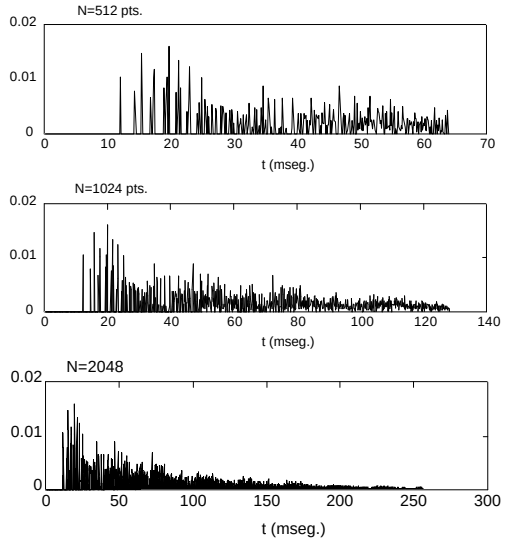

Figure 4: The transference function between any two points of a room is estimated for different functions lengths (64, 128 and 256 miliseconds) with the new room acoustics simulation software

Several options are now available in the system, as correlation or coherence based time delay estimation, delay and sum beamforming, Griffiths-Jim or switched Griffiths-Jim structures [2], and others are being incorporated, as processing based in the decomposition in minimum-phase and all-pass components [8].

4. CONCLUSION

In this paper we have described a robust speaker recognition system working in noisy and reverberant conditions. The joint use of acoustic array processing, coping with the noisy and reverberant speech, and spectral normalization compensating the filtering effects introduced in the array structure has been shown useful trying to get a robust system. Encouraging results have been obtained, but a new software tool had to be designed to allow extensive experimentation with different algorithms, configurations and acoustic conditions.

The structure and philosophy of this tool is reported in this paper, noting the ability of the system to incorporate new algorithms or reconfigure the system to perform new experiments.

ACKNOWLEDGEMENTS

We thank J. Artero for helping us in the incorporation of new algorithms and the development of the ProcAr system.

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

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

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

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

[6] 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.

[7] G.C. Carter, *Coherence and time delay estimation*, Proc. IEEE, vol.73, Feb. 1987, pp. 236-255.

[8] Q. Liu et al., *A microphone array processing technique for speech enhancement in a reverberant space*, Speech Communication, Vol. 18 (1996), pp. 317-334.

[9] S.J. Young et al., *HTK: Hidden Markov Model Toolkit V1.5*, Entropic Research Laboratory, Inc., Dec. 1993.

[10] J. Gonzalez-Rodriguez, J. Ortega-Garcia et al., *Increasing Robustness in GMM Speaker Recognition Systems for Noisy and Reverberant Speech with Low Complexity Microphones Arrays*, Proc. ICSLP'96, pp. 1333-1336, Philadelphia, October 1996.