A New Approach to Dereverberation and Noise Reduction with Microphone Arrays

J.L. Sánchez-Bote, J. González-Rodríguez, and J. Ortega-García
Speech and Signal Processing Group (ATVS)
Departamento de Ingeniería Audiovisual y Comunicaciones (DIAC)
E.U.I.T. Telecomunicación - Universidad Politécnica de Madrid
Ctra. Valencia, km. 7 - Campus Sur, 28031 Madrid
email: jbote@diac.upm.es
http://www.atvs.diac.upm.es

ABSTRACT

In this paper the speech enhancement abilities of a new array-based processor have been tested. The proposed system works in three cascade stages. First, the signals are time aligned with the estimated direction of the desired sound source. Second, the signal is decomposed in its all-pass and minimum-phase components using cepstral processing. In this moment, beamforming and liftering in cepstral domain is performed, where the output signal reverberation is reduced. The third part consists in a noise canceller, based in the optimum Wiener filtering and Coherence evaluation. The processor have been tested both with a real database and in simulated conditions of noise and reverberation. To evaluate its performance, we have considered the subjective perception of the output signal and objective measurements of LAR distances, Real Cepstrum distances and Signal to Noise Ratios.

1 INTRODUCTION

Speech signal acquisition in adverse sound ambients has been an important research field in last years. The main goal of a speech enhancement system is to reduce noise and reverberation to obtain good communication between speaker and receiver. For many years, microphone arrays have been used to achieve this signal enhancement, by means of spatial filtering or beamforming [1][2]. Beamforming consists in aiming the receptor directive beam to the sound source, reducing reverberation and noise, whose origins are, in general, different. Using this method, the improvement is noticeable if the number of microphones is high and if no much reverberation is present. In order to obtain a further reduction in diffuse noise, Wiener post-filtering have been proposed [5]. Also, coherent noise components can be cancelled in our system modifying this filter, as shown in [6]. To obtain better reverberation reduction a system based in All-Pass and Minimum-Phase decomposition, including cepstral liftering, is proposed [3][4].

2 DEREVERBERATION AND NOISE REDUCTION

2.1 Dereverberation using All-Pass and Minimum-Phase processing

The reverberation process in a room with reverberation can be represented by its Impulse Response h(t). This h(t) is like a multipath system response and can be decomposed into its minimum-phase and all-pass components. The two components are written in frequency domain :

$$H(\omega) = H_{min}(\omega) \cdot H_{all}(\omega) \quad (1)$$

To obtain this decomposition is necessary to work in cepstral domain. The reason is that the minimum-phase component can be calculated using the Real Cepstrum of the impulse response.

For a typical h(t), the minimum-phase component has shorter duration than the all-pass component and then it is less disturbed by reverberation. Processing the speech signal at the recording microphones, the same conclusion can be obtained: the minimum phase component of the output speech signal is shorter in time than its all pass component. But now, we have another advantage. The Real Cepstrum of speech signal is, in general, shorter than the same signal with reverberation. So, we can lifter the Real Cepstrum to eliminate only reverberation. However, to achieve enough reverberation reduction it is necessary to cut out excessive cepstrum components, degrading the original signal. It is at this moment when the array processing is useful. Excess reverberation can be eliminated, without excessive liftering, performing spatial filtering (beamforming) in cepstral domain. The system tested performs this beamforming to both minimum-phase and all-pass components. However, some extra processing is necessary, because the beamforming produces some degradation in the all-pass components of the signal.

2.2 Noise reduction using Wiener Filter with Coherence modification

In order to reduce noise contamination, an optimal Wiener filter have been incorporated at the dereverberator output. All signal estimators associated with optimal Wiener filter are extracted by means of extra information coming from the several channels of the system.

The implemented Wiener filter is:

$$H_W(\omega) = \frac{\langle G_{xi\,xj}(\omega)\rangle - \langle G_{ninj}(\omega)\rangle}{G_{xx}(\omega)} \quad (2)$$

where $\langle G_{xi\,xj}(\omega)\rangle$ is the averaged cross spectrum over all channel pairs. Non-coherent noise present in array channels is cancelled by this term because all channels are previously time aligned. Thus, coherent components (desired signal and coherent noise) are not affected by the filter, and diffuse noise (non-coherent part) is cancelled. The term $\langle G_{ni\,nj}(\omega)\rangle$ in (2) represents the averaged cross spectrum over all channel pairs, but considering just temporal frames without speech activity. Consequently, it has the consideration of noise. This noise represents coherent noise (with interchannel coherence) picked up by the array. The numerator of (2) is an estimation of the signal without noise, including coherent and non-coherent noise.

The denominator $G_{xx}(\omega)$ represents the estimation of signal plus noise autospectrum, and it is obtained beamforming all channels in three frequency subbands.

When interchannel coherence is small, the estimations present in (2) can produce wrong synthesis of Wiener filter, and the attenuation of low coherence frequency components is better. The interchannel coherence can be estimated by the expression:

$$C(\omega) = \frac{G_{x0}(\omega)}{\sqrt{G_{xx}(\omega)\cdot G_{00}(\omega)}} \quad (3)$$

where $G_{x0}(\omega)$ is the cross spectrum between beamformed signal and central array channel or reference channel. The term $G_{00}(\omega)$ is autospectrum of reference channel and $G_{xx}(\omega)$ is autospectrum of beamformed signal.

When Coherence C(ω) underpasses a prefixed threshold (coherence threshold), the low coherence frequency component is filtered with the factor $C^{\alpha}(\omega)$.

3 SYSTEM DESCRIPTION

The proposed Array processor, tested in our experiments, is shown in figure 1.

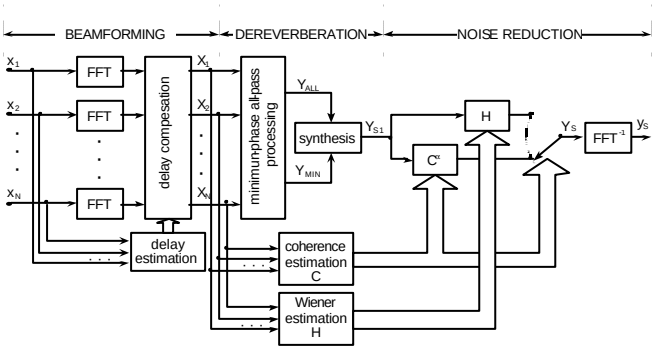

figure 1. Array Processor diagram in three blocks

Each channel input is digital speech signal, sampled at 16kHz, with 16bits quantized samples.

Each one of the system blocks are explained below.

3.1 Beamforming

The Array aiming to the source is estimated by time alignment of all channel signals. To make this time alignment a Time Delay Estimation (TDE) have been used by means of Cross-Power Spectrum Phase method [2]. The estimated delay is applied to the temporal digital speech signal. Because temporal precision of input signals sample period is not enough to obtain correct alignment, it is necessary to interpolate the data to achieve sufficient temporal precision. After all channels are correctly aligned, the signals are decimated by the same factor that the previous interpolation. In our tests, we have used an interpolation ratio between 5 and 15.

3.2 Dereverberation

Time aligned inputs are windowed in frames of sufficient length to treat conveniently the reverberation. Choosing of optimal frame length is not trivial, and depends on reverberation amount and reverberation time, RT_{60}. In our experiments $RT_{60}\approx 1$s, and a 300ms (L=4096 points) window length has been used (Hanning type). Then, all processing detailed in point 2.1 is implemented.

3.3 Noise reduction

Input signals, previously time aligned, have been segmented in 150ms (L=2048 points) frames (Hanning type window). The window length is different to that of dereverberation processing, and thus is necessary to make framing in two stages. Next, all processing detailed in point 2.2 is implemented. A coherence threshold α=50 have been used in our experiments.

4 MULTICHANNEL DATABASES

4.1 Simulated database

This database consists in speech files with Reverberation and Noise artificially added

The model for these experiments is shown in figure 2.

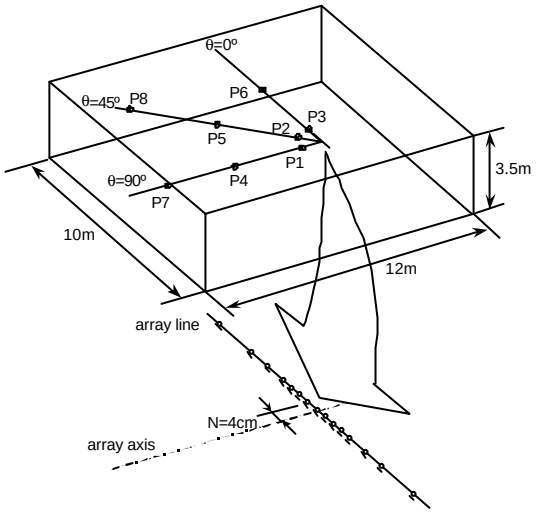

figure 2. Artificial Reverberation and Noise production

Reverberation time, artificially introduced for the tests, was RT_{60}=1s. The array centre is located at 2 meter over the

floor, and with its array line parallel to the rear wall. The array arrangement is shown in figure 3.
Eight test points in the room have been selected (figure 2). These points can take the role of signal or noise sources, according to the type of experiment. Both signal (speech) and noise have been convolved with the Transfer Functions H_T between each probe and each Array microphone.
In this simulated experiments we have taken the next bands for frequency decomposition:

LF=20Hz-1kHz, MF=1kHz-2kHz, HF=2kHz-8kHz

Although high frequencies in LF band have some spatial aliasing, this configuration has been chosen because it has better rejection of reverberation below 1kHz.

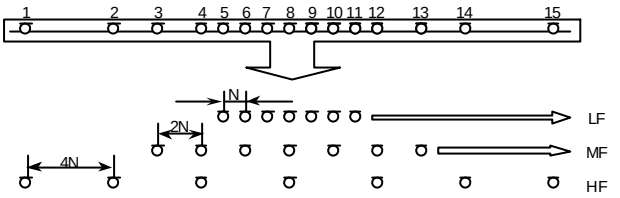

figure 3. Array model and subband decomposition

Two noise types have been added to the signal, according to the experiment. “Coherent noise” is a triangular signal of 500Hz, and “diffuse noise” is a random noise. In all cases the signal to noise ratio introduced in array place was SNR=20dB.

4.2 Real database

Now, input signals consists in Speech files from a Real Database, the Carnegie Mellon University real multichannel database, described in [4], with simultaneous recordings of a reference signal from a head-mounted close-talk microphone and a 15 microphone Array. The Array has the same configuration of figure 3.
The database contains 10 subcorpora, with speech records taken in different rooms and ambient conditions. In this paper we have used the following:
arr3A: Noisy laboratory with low reverberation. Speech source at 1m in array axis.
arrC1A: Meeting Room with high reverberation and noiseless. Speech source at 1m in array axis.
arrC3A: Meeting Room with high reverberation and noiseless. Speech source at 3m in array axis.

5 EXPERIMENTS AND RESULTS

5.1 Tests with artificially added reverberation

The first step to test the array processor have been to study its performance just with reverberation, and no added noise. Experiments with beamforming in three bands, using the nested Array, and beamforming in one band, with no frequency decomposition, have been done. The results, averaged over five different speech utterances are shown in figure 4. Figure 4(a) contains SNR improvements between input and processed signal. SNR is evaluated measuring noise in no speech signal frames. Thus, reverberation is treated here as noise. In figures 4(b) and 4(c) LAR distances and Real Cepstrum distances between original - input signal (with added reverberation) and original - processed signal have been evaluated. The improving is computed comparing input and processed speech with original signal.

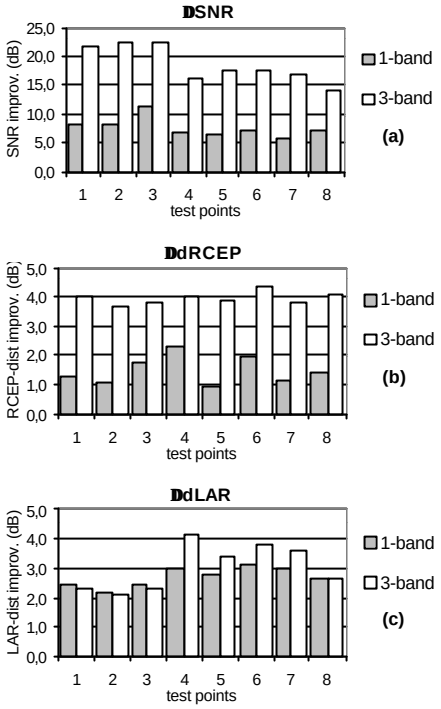

figure 4. Results with just reverberation (no-additive noise). 1-band=no-subband decomposition. 3-band=three frequency subband decomposition. **(a)** SNR improvement. **(b)** Real Cepstrum dist. improvement. **(c)** LAR dist. improvement

Advantages of subband decomposition with regard to whole band processing can be seen in figures 4(a), 4(b) and 4(c). It is clear that the decomposition in three frequency bands have benefits over the simpler one band analysis. In horizontal axis it can be seen the eight points of analysis, numbered in figure 2. With regard to SNR improvement (figure 4a), the system works better when the signal source is close to the array and thus, there is less reverberant acoustic field compared with direct field. We can also observe improvements both in LAR and RCEP distances in any listening point.

5.2 Tests with artificially added reverberation and noise

The next step was to apply noise coming from different points in the room. “Coherent noise” and “Diffuse noise” have been added as described item 4.1. The processor was tested using three-subband decomposition.
Averaged results (over 5 speech files) are shown in fig. 5.
Worst results occur when the noise source is in array axis, and the signal source is out of axis. The little and negative improvements in dLAR measures indicates, that in presence of much noise, this objective parameter can be useless to determine the improvement produced by the Array processor. Nevertheless, in all cases the subjective evaluation of the processor performance was good or very good. The extraction process of objective improvement measures is shown in figure 6.

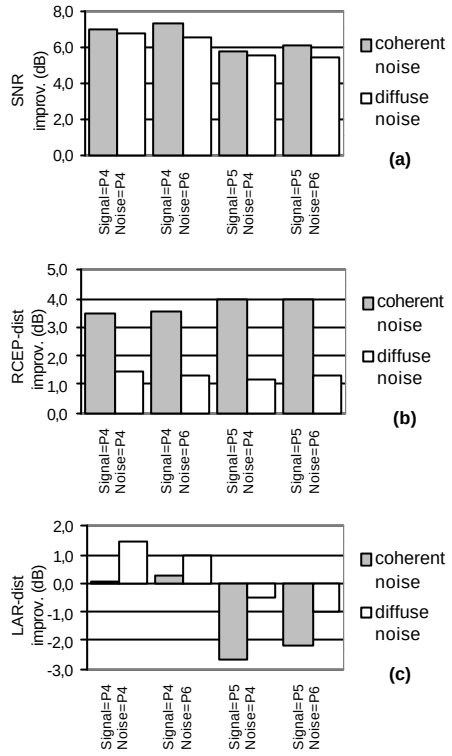

figure 5. Results with reverberation and additive noise in different signal and noise locations (see legend). **(a)** SNR improvement. **(b)** Real Cepstrum distance improvement. **(c)** LAR distance improvement

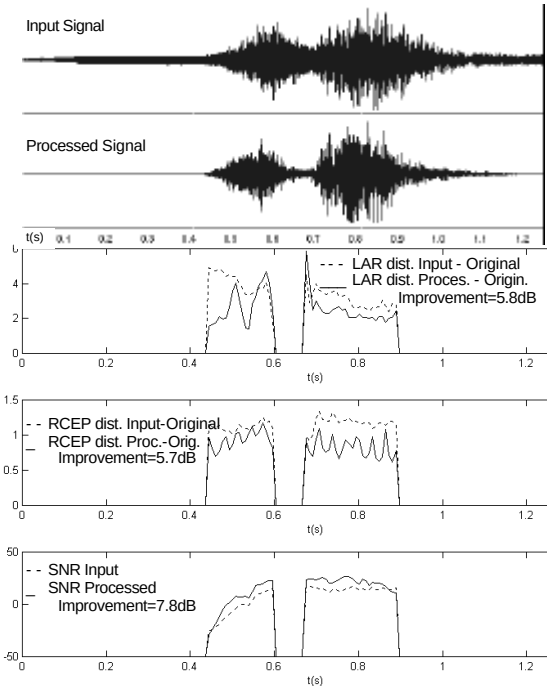

figure 6. Sample of objective improvement measures

5.3 Tests with Real Database

When the Array processor was tested using the real database, LAR and RCEP measures have been proven bad indicators of achieved improvements. In figure 7 SNR improvements, averaged over five tests can be seen. Results are better when the room is no much reverberant, and has important noise.

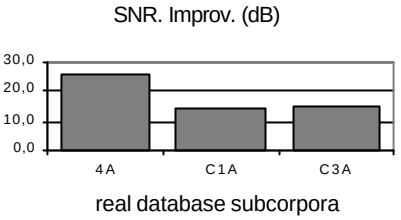

figure 7. Results with SNR improvement for real database

6 CONCLUSIONS AND FUTURE WORK

Subjective evaluation of the processed speech signals was pretty good. These signals have shown less reverberation and much less noise.

Objective evaluation of the results was good in general, and very good when just reverberation is added to the original speech signal.

When speech signal is highly contaminated by noise, LAR distances and RCEP are erratic, and in some case no improvements are achieved, although the subjective evaluation of the results was good.

In future we will investigate another objective evaluators like RASTI or other intelligibility estimators, to achieve better correspondence with subjective tests.

References

[1] Van Been, B.D., and Buckley, K.M., "Beamforming: a Versatile Approach to Spatial Filtering", *IEEE ASSP Magazine*, April 1988, 4-24 (1988).

[2] M. Omologo and P. Svaizer, "Use of the Cross-Power Spectrum Phase in Acoustic Event Localization", *IEEE Trans. on Speech and Audio Processing*, vol. 5, No. 3, pp. 288-292, September 1997.

[3] Q.G. Liu, B. Champagne and P. Kabal, "A Microphone Array Processing Technique for Speech Enhancement in a Reverberant Space", *Speech Communication*, vol. 18, pp. 317-334, 1996.

[4] J. Gonzalez-Rodriguez et al., "Speech enhancement with microphone arrays through minimum-phase and all-pass decomposition" (in Spanish), Proc. of TecniAcustica'99, Avila (Spain), October 1999.

[5] Zelinski, R., "A Microphone Array with Adaptive Post-filtering for Noise Reduction in Reverberant Rooms", Proc. of ICASSP'88, pp. 2578-2581, 1988.

[6] Le Bouquin-Jeannes, R. et al., "Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator", *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 484-487, Sep. 1997.

[7] J. Gonzalez-Rodriguez et al., "Speech dereverberation and noise reduction with a combined microphone array approach", Proc. of ICASSP 2000 (accepted).