SPEECH DEREVERBERATION AND NOISE REDUCTION WITH A
COMBINED MICROPHONE ARRAY APPROACH

J. Gonzalez-Rodriguez, J.L. Sanchez-Bote and J. Ortega-Garcia

Speech and Signal Processing Group (ATVS)
D.I.A.C.- E.U.I.T. Telecomunicacion - Universidad Politecnica de Madrid
Ctra. Valencia, km. 7 - Campus Sur, 28031 Madrid
email: jgonzalz@diac.upm.es
http://www.atvs.diac.upm.es

ABSTRACT

In this contribution we have addressed the problem of speech enhancement in noisy and reverberant rooms through the use of a new approach that combines the dereverberation abilities of a structure based in the separate processing of the minimum-phase and all-pass components of the input speech signals, and the noise rejection performance of a speech-activity-based Wiener filter able to cope both with coherent and diffuse noise. Experiments have been performed with the CMU real multichannel database, which includes a clean speech reference through a head-mounted microphone. This reference signal has been also used to perform simulation experiments in controlled conditions. Extensive results have been obtained, both with LAR and cepstral distances of input and processed signals to the reference, and with segSNR improvements, assessing the abilities of the new system to cope both with reverberation and coherent and diffuse noise in different acoustic environments.

1. INTRODUCTION

Speech enhancement for dereverberation and noise reduction in reverberant rooms has been extensively addressed. Even though beamforming [1][2] to the speaker and steering zeros in the direction of the noise arrival is a reasonable option when one or several noise sources are present in free field conditions, this approach will not be enough if moderate or strong reverberation is present, caused by severe multipath propagation in rooms. In order to cope with reverberation, the usual beamforming is improved in [3] with the addition of a new processing stage, where the speech input signals are decomposed into its minimum-phase/all-pass components and are separately processed, having shown to be extremely effective for speech dereverberation [4], because reverberation mainly affects the all-pass components, obtaining an almost reverberation-free speech signal from the minimum-phase components.

However, the usual condition in offices or meeting rooms is that moderate reverberation is present together with noise components. These noise sources, as computer fans or air conditioning systems, are placed both near and/or quite apart from the receivers. In this case, we can not talk of a direction of arrival of the noise signal, and a combination of a coherent noise field, associated with the direct path and early reflections, and a diffuse one, associated with the late reflections, will be present together. Recently, the coherent/non-coherent nature of the sound field has been used to separate the wide-band speech signal into coherent/diffuse subbands [5], with promising results for diffuse noise with wavelet-domain processing. However, this system will fail in the presence of coherent noise sources, obtaining similar performance to Wiener postfiltering, as in Zelinski’s proposal [6], because any coherent component is supposed to arrive from the desired signal source.

But we can take advantage of the burst nature of the speech signal, and we can learn the coherent noise components in speech absence, as shown in [7], and modify the Wiener filter to be able to cancel both diffuse and coherent noise components, as has been tested by the authors [8] with excellent results in noise reduction in different conditions.

2. SYSTEM DESCRIPTION

In order to obtain better dereverberation properties, and the same noise rejection abilities of the former author's system [8], we propose in this contribution a processing scheme that works with three cascade stages, performing:

1) Beamforming: time delay estimation (TDE) and time delay compensation (TDC) is performed through the Cross-Power Spectrum Phase [2] method.

2) Dereverberation: decomposition and separate processing of the minimum-phase and all-pass components of the speech input signals [3].

3) Noise reduction: an optimal Wiener post-filter is applied in order to cancel both coherent and diffuse noise [8].

2.1 Speech dereverberation

In order to see this behavior, we show in the following figure the minimum-phase and the all-pass components of a room impulse response, where we can see that:

- The minimum-phase component is less affected by reverberation than the all-pass component. In this way, all information from the original signal is present in the minimum-phase component of the received signals.
- The all-pass component has all the information related to source localization and propagation.
Following these observations, we are going to perform separate processing of the minimum-phase and all-pass components of the input speech signals, as described in [3]. Starting from the fact that the complex cepstrum of the speech signal is concentrated in low quefrencies, and the information about echoes is quite apart in this domain, each one of the array signals is decomposed into its minimum-phase and all-pass components, performing a different processing with each one of them:

- **Minimum-phase processing:** the minimum-phase component, directly related to the real cepstrum, is less affected by reverberation. In this way, in each one of the minimum-phase components of the array we will find the same speech signal and some information about the transference function between source and each one of the microphones, different for different channels. Then, we perform a spatial averaging of these components, enhancing the speech signal. Furthermore, this processing is equivalent to geometric averaging in the Z plane, preserving the minimum-phase sequence property (poles and zeros inside the unit-circle). This averaging process is followed by a low-pass liftering selecting the signal components and rejecting components from reverberation.

- **All-pass processing:** in the all-pass components, phase information of each input channel is preserved, performing then a spatial filtering through frequency adding of each component. The resulting signal is not all-pass, so we have to extract the all-pass component of this signal to further combine it with the minimum-phase component previously computed.

### 2.2 Noise Reduction

Even though dereverberation is of great importance in speech enhancement for far utterances (one to several meters), even more important is the presence of different noise sources in the room. Trying to avoid the residual noise that remains in speech after the described dereverberation process, an optimal Wiener post-filter is designed and applied at the output of the dereverberator. Therefore, if we suppose that the noise components in the different microphones are uncorrelated due to the spatial separation between microphones, we can design and optimal Wiener filter, as shown in [6], given by:

\[
H_{w}(f, k) = \frac{\gamma_{i,i}(f, k)}{\gamma_{i,i}(f, k)}
\]

where in the numerator we have the average of the cross-spectral densities of the input signals in frame k, and in the denominator the average of the auto-spectral densities of the input channels.
3. SYSTEM EVALUATION

3.1 Multichannel Database

In this work, we have used a multichannel database recorded by T.M. Sullivan and R.M. Stern from Carnegie Mellon University (CMU), Pittsburg (PA, USA). This database contains simultaneous recordings of clean speech, through a head-mounted close-talking microphone, and multichannel recording from the microphone array, which gives us an exact reference of the effects introduced by the acoustic propagation process. The database, sampled at 16 KHz, contains several sub-corpora, as described in [8], each one of them with 14 different utterances. The subcorpora we have used were recorded by the same speaker with a 15-element array spaced in order to have available three 7-element sub-arrays interleaved, with linear spacing of N, 2N and 4N respectively, as we can see in figure 3.

Figure 3.- Recording microphone array configuration with correspondent spacing of N, 2N and 4N for each seven microphone subarray

With the minimum spacing of N=4 cm. that we have chosen, each one of the 4N, 2N and N subarrays will cover the subbands of 0-1 kHz, 1-2 kHz and 2-8 kHz respectively. The used subcorpora are the following:

- **arr4A**: noisy lab, minimum spacing (N) of 4 cm., subject sat one meter from the array (d=1 m.)
- **arrC1A**: collected in a conference room, larger than the noisy lab but much more quiet, N=4 cm. and d=1m.
- **arrC3A**: same conference room, N=4 cm., but now d=3 m.

In spite of the great interest in the evaluation of the proposed array processing structure with this database, we have in real situations, as it happens in this database, all the undesired effects together (reverberation and noise). Then, it is also interesting to test the system in controlled situations, and with this objective we have developed from the clean reference signals of this database, recorded with the head-mounted microphone, a simulated database called SimCMU, where we have available:

- Array recordings with different reverberation times (T60=0.3, 0.5 and 0.8 seconds)
- Different distances to the speaker (d=1 and 3 meters)
- Fixed and moving speakers, with known trajectories, in order to evaluate our target tracking abilities
- Coherent and diffuse noise (both individually and joint) with several input SNR.

SimCMU have been obtained from convolution of the clean reference signals from CMU with the room impulse response computed in each configuration from the speaker position to each one of the microphones of the array, obtaining the transfer functions with the image method in controled rooms (wall absorption coefficients, position of speaker and array, room size) [8]. The combination of the different characteristics shown above give us a total of about 3 Gbytes of available information in SimCMU, compared to the 300 Mbytes of the original CMU database.

3.2 Experiments and Results

The whole CMU subcorpora just described have been processed with the proposed structure, obtaining excellent results both in perceived quality and SNR improvements. Several audio examples (reference, input and processed signals in `wav` format), both real (CMU) and simulated (SimCMU) can be seen and heard from our web page at http://www.atvs.diac.upm.es. As an example, we show in the following figure the input and processed signals for a file from SimCMU with loud diffuse noise and rather high (T60=0.8 s., d=3 meters) reverberation:

Figure 4.- Example of input and processed signals with the proposed structure.

The objective evaluation of the performance of the proposed system with these data have been done with the following parameters per processed file, and later averaging over the 14 utterances of each subcorpora:

- **SNR (dB) improvement**, computed as SNR\text{out} - SNR\text{in}.
- **Average LAR (Log Area Ratio) distance improvement**, computed as dLAR\text{input-ref} - dLAR\text{output-ref} , where each frame distance is obtained as euclidean distance between LAR parameters of each frame.
- **Average cepstral distance (dCEP)**, obtained as dLAR but computing real cepstral parameters.

SNR is computed without human supervision for each file with a double threshold strategy, where frames over the higher threshold are taken as signal and frames under the lower one are taken as noise, and computing the 10log of the average energy of signal and noise frames.
In the following figure we show an example of dLAR and dCEP computation per frame, and the mean of the improvement for the processed file:

![Graph showing dLAR and dCEP computation](image)

Figure 4.- Example of input and processed signals with the proposed structure.

In the following tables, we sum up the average results over 14 utterances for each one of the tested subcorpora in different conditions, obtaining objective improvements in all cases:

<table>
<thead>
<tr>
<th>Improvement</th>
<th>$T_{60}$ (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1 s.</td>
</tr>
</tbody>
</table>

$\Delta$SNR (dB) | 19.2 | 12.4 | 10.0 |

$\Delta$dLAR | 0.94 | 0.83 | 0.73 |

$\Delta$dCEP | 0.14 | 0.11 | 0.10 |

Table 1.- SNR, dLAR and dCEP improvements for an input signal with SNR$_{in}=10$ dB, for different reverberation times.

<table>
<thead>
<tr>
<th>Improvement</th>
<th>SNR$_{input}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 dB</td>
</tr>
</tbody>
</table>

$\Delta$SNR (dB) | 10.0 | 6.6 |

$\Delta$dLAR | 0.7 | 0.3 |

$\Delta$dCEP | 0.1 | 0.4 |

Table 2.- SNR, dLAR and dCEP improvements for an input signal recorded with $T_{60}=0.8$ s., for different input SNR.

In this paper we have taken advantage of the dereverberation abilities of the separate processing of the minimum-phase and the all-pass components of the input speech signals, and have canceled the residual noise, with coherent and/or diffuse nature, through the use of a voice activity based modified Wiener postfilter. The system has been extensively tested in real and simulated environments, obtaining excellent results in both objective measurements, and subjective listening tests (audio files available at http://www.atvs.diac.upm.es).

4. CONCLUSION

In this paper we have taken advantage of the dereverberation abilities of the separate processing of the minimum-phase and the all-pass components of the input speech signals, and have canceled the residual noise, with coherent and/or diffuse nature, through the use of a voice activity based modified Wiener postfilter. The system has been extensively tested in real and simulated environments, obtaining excellent results in both objective measurements, and subjective listening tests (audio files available at http://www.atvs.diac.upm.es).

5. REFERENCES


