COHERENCE-BASED SUBBAND DECOMPOSITION FOR ROBUST SPEECH AND SPEAKER RECOGNITION IN NOISY AND REVERBERANT ROOMS

Joaquin Gonzalez-Rodriguez, Santiago Cruz-Llanas and Javier Ortega-Garcia

Dpto. de Ingenieria Audiovisual y Comunicaciones
EUIT. Telecomunicacion, Universidad Politecnica de Madrid, SPAIN
e-mail: jgonzalz@diac.upm.es

ABSTRACT[*]

In this paper, the acoustic characteristics of sound fields in enclosed rooms are studied in the joint presence of speech and noise, in order to design a broadband microphone array system capable of coping with both coherent and diffuse noises. Several state-of-the-art speech enhancement array structures are presented and compared to our new system in terms of correct word recognition rates in a simple command and control task. The proposed structure, based on a broadband subband-nested array, performs real-time estimations of the spatial coherence in order to determine the coherent/diffuse nature of the different subbands, using different filters in each case, improving also the classical Wiener post-filter, typically used for diffuse noise supression, for proper cancellation of coherent noises. The results obtained with a 15-channel simultaneous recording database in different reverberation and noise conditions show better performance than other structures previously proposed.

1. INTRODUCTION

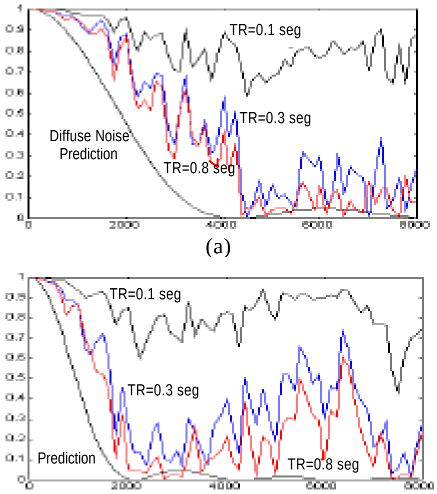
Speech enhancement for robust recognition in reverberant rooms has been extensively addressed. Though beamforming towards the target speaker and steering zeros in the direction of the noise arrival is a reasonable approach when one or several noise sources are present in free field conditions, this approach will fail if moderate or strong reverberation is present. A modified Griffiths-Jim structure was tested by the authors as pre-processing stage to a speaker identification system with excellent results for a single noise source quite close to the microphone array [8]. However, the usual condition in offices or meeting rooms is that moderate reverberation is present and the 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 coherence/non-coherence 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 as Zelinski's proposal [3]. A combined structure with Generalized Sidelobe Cancelling for coherent bands [2] and Wiener filtering for the diffuse ones [3] has been tested previously by the authors [9] with excellent results at high computational cost. The processing scheme proposed here takes advantage of a three subband nested array in order to avoid spatial aliasing, while further processing is performed based on real-time estimations of spectral coherence and speech activity detection, allowing excellent cancellation of both diffuse and coherent noise.

2. SOUND FIELD IN REVERBERANT ROOMS

In diffuse pure-tone sound fields, the phase difference between two pressure signals can be regarded as a random variable. In such a room, the coherence of two pressure signals can be estimated [1] by the expression:

$$g_{pp}^{2}(w,r)=\left[\frac{\sin(w\cdot r/c)}{(w\cdot r/c)}\right]^{2}$$

where *r* is the distance between microphones. In the other hand, as the sound field approaches the free field condition, the coherence approaches to one. Consequently, we can expect an intermediate situation in usual rooms with speech presence as we can observe in figure 1.



(a)

(b)

Figure 1.- Spatial coherence as a function of frequency for different reverberation times in a simulated room with (a) *r* = 4 cm and (b) *r* = 8 cm

[*] This work is supported by CICYT under TIC97-1001-C02-01

3. MULTI-CHANNEL RECORDED SPEECH DATABASE

Speech enhancement systems based on microphone arrays are usually tested with simulated data. This simulation process is performed through the image method [7], based on the acoustic ray theory. However, this method is a valid approximation of the true acoustic propagation process just if the wavelength is at least smaller than the third part of the smallest dimension for an empty room. Then, for usual rooms the method could be used just for frequencies over several hundred Hertzs. Moreover, the designed systems are not intended for empty rooms but for everyday ‘full-equipment’ rooms, with lots of different elements (furniture, computers...) much smaller than the room dimensions, reducing the usable frequency range at least to frequencies over 1 KHz. If we work with speech data, these limitations of the model are added to the fact that the frequency content of speech is mostly concentrated (energy, spectral formants, pitch) below 1 KHz. Therefore, as we do not have available a simple and efficient method to simulate actual propagation in rooms, we are forced to use actual multichannel speech data in order to evaluate our systems as the complexity of the algorithms becomes higher and can be suited for the simulated data and not able to cope with the acoustically propagated data in real situations.

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, namely:

- *arrA*: 10 male speakers speaking at a distance of 1 meter from the center of a 7 cm. 8-elements linearly spaced array in a noisy lab with many computers and disk-drive fans.

The remaining 6 sets 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 2.

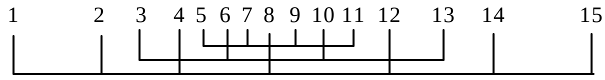


Figure 2- Recording microphone array configuration with correspondent spacing of N, 2N and 4N for each seven microphone subarray

- *arr3A*: same noisy lab as above, minimum spacing (N) of 3 cm., subject sat one meter from the array (d=1 m.)
- *arr4A*: noisy lab, N=4 cm., 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.
- *arrCR1A*: same conference room, N=4 cm., d=1 m., but including an AM talk-radio jamming signal at approximately 45 degrees off-axis from the center of the array, competing with the speaker.
- *arrCR3A*: idem as before, but d=3 m.

In this work we have just used the *arr4* (noisy lab) and *arrC1A/arrC3A* (conference room, d=1m./d=3m.) as we will have to deal and evaluate the influence of different distances in the same conference room, and the effect of strong coherent noise in the noisy lab. 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.

We have computed for the noisy lab the spatial coherence for the different separations between microphones, and we can observe in figure 3 the joint presence of the diffuse noise field and strong coherent noise contributions.

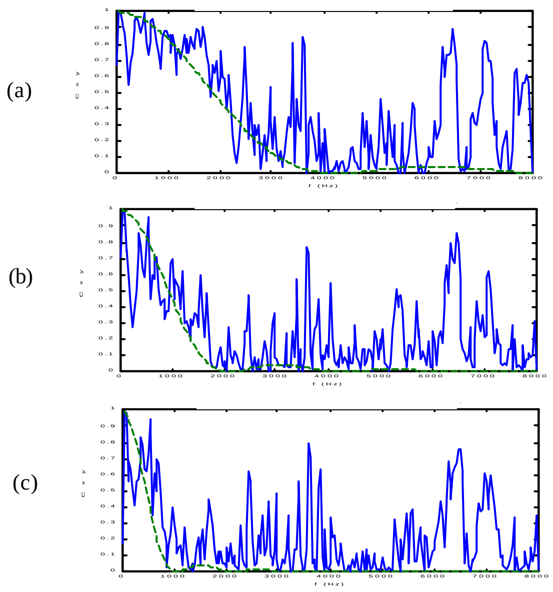


Figure 3- Spatial coherence measured in the noisy lab and diffuse noise prediction as a function of frequency for different separation of microphones (a) r = 16 cm., (b) r = 8 cm. and (c) r = 4 cm.

This joint presence of coherent and diffuse noise will be an usual situation in application rooms, so our speech enhancement algorithms will have to be able to cope together with both types of noises.

We can also see in figure 4 what we could call a ‘coherencegram’ from the *arrC3A* subcorpus. A coherencegram would be then a representation of the spatial coherence, averaged in 500 Hz subbands in this case, for each time frame, where the spatial coherence for the whole frequency band has been obtained from the partial estimations of the different subarrays. It is shown together with the spectrogram of one of the input signals from the microphone array in order to avoid confusions from the graphical information provided. In this way, we can observe background noise just for frequencies below 500 Hz. However, we can see in the coherencegram strong coherence up to 2000 Hz due to background noise and microphone spacing.

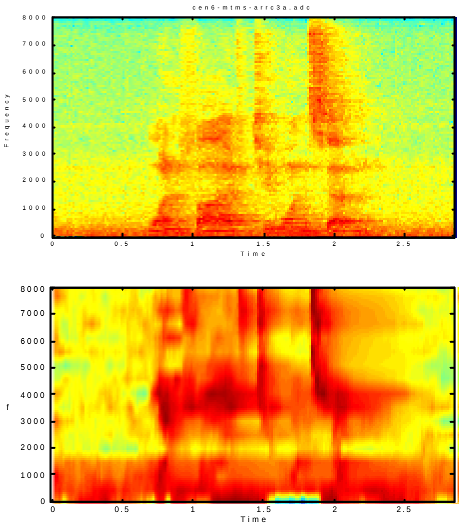


Figure 4.- Spectrogram of one of the input signals and 'coherencegram' in the conference room (*arrC3A*)

4. COHERENCE-BASED PROCESSING

In order to evaluate the performance of the structure that we are going to propose, several systems are going to be implemented:

- Conventional beamformer (*BF*): ideal time delay compensation is performed in the frequency domain, obtaining a directivity gain which will also be used in all of the following systems.

- Wiener post-filtering (*W*)[3][4]: based on the assumption of spatially uncorrelated noise, the Wiener postfilter in frame *k* is estimated from:

$$H_W(f,k) = \frac{g_{ss}(f,k)}{g_{xx}(f,k)}$$

  where the input *x* is composed of the clean signal *s* and the noise *n*. The presence of correlated noise (coherent noise) will introduce errors in the estimation of the filter which will affect the quality of the processed signal.

- Coherence based processing (*CbP*)[5]: further reduction of non-coherent noise through a coherence-shaped filter in the low coherence subbands. In detail, after computing the coherence between the input and the beamformed signal:

$$C_{x,x_{BF}}(f,k) = \frac{\left|g_{x,x_{BF}}(f,k)\right|}{\sqrt{g_{x,x}(f,k)\cdot g_{x_{BF},x_{BF}}(f,k)}}$$

  these frequency dependent values are averaged in subbands, applying the following processing in the m^{th} subband:

$$\begin{aligned} &if \quad C_m > T \Rightarrow H_m(f,k) = H_W(f,k) \\ &if \quad C_m < T \Rightarrow H_m(f,k) = C_m(f,k)^a \end{aligned}$$

  The inconvenient of this approach is that in the case of coherent noise, the corresponding subband will be marked as coherent and it will be unafected by the system.

- Le Bouquin and Faucon system (*LBF*)[6]: as the coherent components are going to be derived from speech and coherent noise, by means of a speech activity detector we can learn the spectral characteristics of the coherent noise in speech absence, and modify the Wiener filter to cancel its effects:

$$\hat{H}_W(f,k) = \frac{\hat{g}_{ss}(f,k)}{g_{xx}(f,k)} = \frac{g_x(f,k) - g_n(f,k)}{g_{xx}(f,k)}$$

  This system lacks of the advantage over diffuse noise taken into account in the former approach.

- Modified Coherence-based Processing (MCbP): this is the system we are proposing in this paper, and takes advantage of the different ideas previously exposed, in order to be able to cope with both coherent and diffuse noise. In this way, the system performs the following:

$$\begin{aligned} &if \quad C_m > T \Rightarrow H_m(f,k) = \hat{H}_W(f,k) \\ &if \quad C_m < T \Rightarrow H_m(f,k) = C_m(f,k)^a \end{aligned}$$

5. EXPERIMENTS AND RESULTS

These systems has been tested with real data from the CMU-multichannel database, using the *arr4A* (noisy lab) and *arrC1A/arrC3A* (conference room) subcorpora, in order to evaluate its performance under reverberant and noisy (both coherent and diffuse) speech.

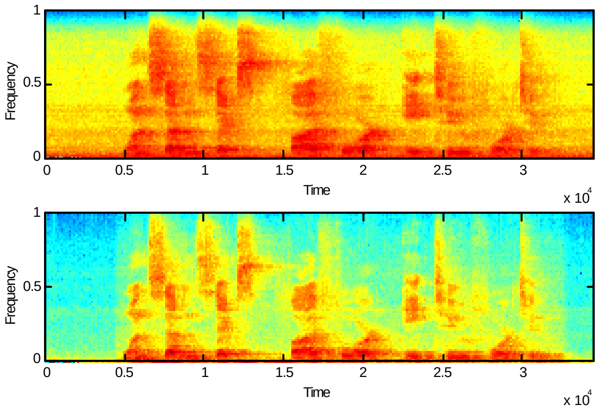


Figure 5.- Spectrograms of the input signal (with noise and reverberation) and the processed signal through the proposed structure

First of all, the systems has been optimized for speech enhancement, obtaining the best configuration parameters in order to get the best perceptual quality. In figure 5, we can observe the spectrograms of a reverberant signal from the conference room, and the processed signal through the proposed structure, where both the background noise and the reverberation are severely reduced with no audible distortion.

Secondly, these structures has been tested as preprocessing stage to a command and control word recognition system. We have trained simple phone models from the Resource Management database, and constructed a word grammar, where digits, numbers, commands and spelled letters can be uttered in an isolated or continous way. The baseline system has been obtained testing with the close-talking signals from the database with around 80% of correct word recognition.

These, of course, are the highest rates that we coul obtain in the case of perfect reconstruction of a 'clean' signal. In table 1 we can observe the recognition results for the different evaluated systems and the different subcorpora tested.

	Ref.	Mic.	BF	W	CbP	LBF	MCbP
arrC1A	80.9	63.6	65.5	70.0	66.4	68.2	68.2
arrC3A	82.7	52.7	54.5	58.2	58.2	59.1	59.1
arr4A	79.1	48.2	50.0	52.7	40.9	47.3	57.3

Table 1.- Correct Word Recognition rates with different subcorpora with different processing systems

As we can observe from table 1, the effects of reverberation and noise are reflected in great reductions in recognition rates when no further processing is applied (*Mic.*), increasing with the distance from the array and the noise level. From the table, it is clear than conventional beamforming (*BF*) has little effect over recognition rates, while Wiener filtering (*W*) of the diffuse noise contributions are noticeable for the conference room (basically diffuse noise) but smaller for the noisy lab (strong coherent and diffuse noise).

With respect to the original coherence based processing (*CbP*), it obtains better perceptual results than the Wiener filtering approach when no coherent noise is present. However, the nonlinearities introduced makes the recognition system to have worse or equal results than Wiener ones. But in the case of the noisy lab, due to the presence of strong coherent noise, the recognition results drop dramatically due to considering that all coherent contributions come from the desired signal, and then passing them almost unaffected. The *LBF* system, however, obtains the better rates for the conference room taking into account the coherent nature of low frequency noises, but does not take extra advantage of the diffuse noise also present in the noisy lab.

Finally, and as it was expected, the best results are those obtained by the proposed system (*MCbP*), who equals the results obtained by *LBF* in the conference room, but clearly outperforms the other systems when strong coherent and diffuse noise are present together, as it is the case of the noisy lab.

6. CONCLUSIONS

After an in depth analysis of the characteristics of the sound field in enclosed rooms, a new system has been proposed which is able to cope both with diffuse and coherent noise. The obtained results from a real multichannel database in different situations confirm the theoretical expectations, converting the system in an excellent, low computational cost alternative for any room or new acoustic environment due to the self-adjusting characteristics of the proposed system

7. ACKNOWLEDGEMENTS

The authors wish to thank Dr. Finn Jacobsen (Technical Univ. Denmark) for his fruitful discussions about spatial coherence, and Dr. Richard Stern (CMU) for the multichannel database.

8. REFERENCES

1. Jacobsen, F., and Nielsen, T.G., "Spatial Correlation and Coherence in a Reverberant Sound Field", *Journal of Sound and Vibration*, 118, 175-180 ,1987.

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

3. Zelinski, R., "A Microphone Array with Adaptive Post-filtering for Noise Reduction in Reverberant Rooms", *Proc. of the IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2578-2581, 1988.

4. S. Fischer and K. Kammeyer, "Broadband Beamforming with Adaptive Postfiltering for Speech Acquisition in Noisy Environments", *Proc. of ICASSP'97*, Munich (Germany), 359-362, 1997.

5. D. Mahmoudi and Andrzej Drygajlo, "Combined Wiener and Coherence Filtering in Wavelet Domain for Microphone Array Speech Enhancement", *Proc. of ICASSP'98*, Seattle (USA), 385-388, 1998.

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, September 1997.

7. 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.

8. Gonzalez-Rodriguez, J. and Ortega-Garcia, J., "Robust Speaker Recognition through Acoustic Array Processing and Spectral Normalization", *Proc. of ICASSP'97*, Munich (Germany), 1103-1106, 1997.

9. Gonzalez-Rodriguez, J. and Ortega-Garcia, J., "Coherence-based Subband Decomposition for Efficient Reverberation and Noise Removal in Enclosed Sound Fields", *Proc. of the ASA International Conference On Acoustics (ICA'98)*, Seattle (USA), 1998.