Speech enhancement and recognition by integrating adaptive beamforming and Wiener filtering

Alberto Abad and Javier Hernando

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{alberto, javier}@talp.upc.es

Abstract

A robust adaptive beamforming method is presented in this paper for speech enhancement and speech recognition with microphone arrays. The proposal is based on a modification of the Generalized Sidelobe Canceller with adaptive blocking matrix and the use of a Wiener filter. Alternatively to most of the previous reported works based on microphone arrays with postfiltering, the new technique integrates the Wiener filter in the structure of the adaptive beamformer in a single stage. Experimental results show that the proposed integrated adaptive Wiener-filtering (IAW) beamformer usually is more robust to directional and ambient noises than conventional postfiltering of the beamformer output with a lower level of degradation. Speech recognition experiments which show improvements with the proposed beamformer are also reported.

1. Introduction

Some challenging speech applications, such as video-conferences and smart-rooms, might use microphones that can be several meters away from speakers. In these conditions recorded signals are severely degraded by noise and reverberation, and usually some kind of processing is necessary to enhance the speech signal.

One field of growing interest to reduce problems introduced by distant microphone recordings is multimicrophone processing. More concretely, microphone array processing [1, 2] has been broadly used as a pre-processing stage in order to enhance the recorded signal that might be used for any speech application, particularly, for speech recognition.

Many different proposals exist for microphone array designs but most of them can be basically summarized into two major trends: fixed and adaptive beamforming. On one hand, fixed beamformers as the Delay&Sum (DS) [1] are quite simple solutions but are limited by the number of microphones and unable of reducing highly directive noise sources. On the other hand, adaptive beamforming, like Generalized Sidelobe Canceller (GSC) [3] based techniques, present a higher capability of interference cancellation but they are much more sensitive to steering errors and suffer from signal leakage and degradation.

In order to overcome some of the drawbacks of fixed and adaptive beamforming different robust solutions are used. A postprocessing Wiener filtering stage is usually applied to the output of beamformers to improve the performance for diffuse noise fields [4]. To solve the problems of the adaptive beamforming, Hoshuyama et al. [5] propose using an adaptive blocking matrix (ABM) where coefficients are constrained to allow a determinate target error region.

In this work we integrate adaptive GSC-like beamformer with ABM and Wiener filtering by modifying the fixed beamformer (FBF) part. We apply Wiener filtering to generate a cleaner output of the FBF that is used to design the ABM and therefore, to generate the noise references for the multiple-input canceller (MC) part of beamformer. Thanks to this integration we expect to get a higher directive and diffuse noise cancellation with less distortion of the speech signal in a single stage.

In section 2 and 3 we shortly overview the GSC based scheme used in this work and the Wiener postfiltering. In section 4 the proposed integrated adaptive Wiener-filtering (IAW) beamformer is described and some expected advantages are shown. Several speech enhancement and speech recognition experimental results that corroborate usefulness of the new proposal are finally shown in section 5.

2. Robust adaptive beamforming based on a GSC structure

A GSC beamformer basically consists of a fixed and an adaptive path. The adaptive path tries to estimate the non-desired components through a spatial blocking matrix that blocks target signal direction and allows all the other directions. These non-desired components are used for reducing the correlated components of the output of the fixed beamformer in order to obtain a cleaner output. Usually, most common robust adaptive beamforming techniques are modifications of this GSC structure designed to reduce target signal cancellation.

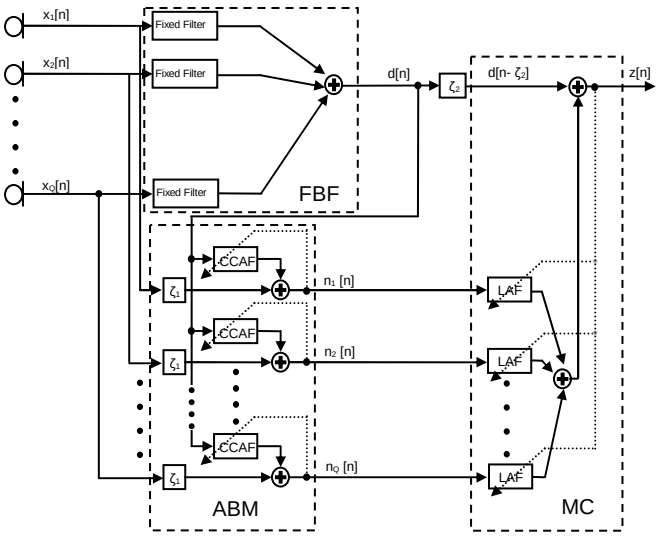

Figure 1. GSC with a CCAF-LAF structure.

The beamformer proposed by Hoshuyama et al. in [5] is named CCAF-LAF (coefficient constrained adaptive filters and leaky adaptive filters) structure and can be seen in Fig.1. It is a robust modification of the GSC, where the blocking matrix (BM) is adaptively designed to allow a concrete target-looking error region and to minimize the leakage of the desired signal at the beamformer output.

The CCAFs in the adaptive blocking matrix (ABM) minimize the output of the ABM resulting in target tracking. Thus, CCAFs processing consists on a NLMS algorithm constrained to a maximum and a minimum bound for the values of the coefficients, according to the maximum allowable look-direction error desired.

Leaky adaptive filters (LAFs) are used in the multiple-input canceller (MC) to minimize the components of the fixed beamformer output (FBF) correlated with the outputs of the ABM enhancing the robustness of the system. It also consists, on a NLMS updating process, with a small leakage constant that avoids excess growth of tap coefficients and that prevents the signal target cancellation when minimization at the ABM is incomplete.

3. Microphone arrays with postfiltering

Usage of Wiener postfiltering techniques with microphone arrays has been deeply studied in [4] by Marro et al. In that work it is shown that postfiltering the output of a microphone array suppress the uncorrelated components of the signals and enhance the performance of the beamformer.

The underlying idea consists in assuming that noise and reverberation components form a diffuse noise field and therefore they are uncorrelated at each microphone of the array. In that case, Wiener optimal filter can be estimated thanks to the availability of multiple inputs that permits computing the power spectral density of the target signal and the one of the noise combining the cross-power spectral densities and the power spectrum density of the different microphones of the array. The optimal Wiener postfilter for the case of the Delay & Sum beamformer can be written as,

$$W(f)=\frac{\frac{2}{Q(Q-1)}\mathrm{Re}\left[\sum_{i=1}^{Q-1}\sum_{j=i}^{Q}\hat{\Phi}_{v_iv_j}(f)\right]}{\frac{1}{Q}\sum_{i=1}^{Q}\hat{\Phi}_{v_iv_i}(f)} \quad (1)$$

where f is the frequency index, Q is the number of microphones, $\hat{\Phi}_{v_iv_j}$ is the estimated cross-power spectral density between the time compensated microphone signals i and j and $\hat{\Phi}_{v_iv_i}$ the estimated power spectrum density of the delayed signal of the *i-th* microphone (see figure 2).

4. The integrated adaptive Wiener-filtering beamformer

There have been some previous works where a similar idea of using a robust beamformer based on modifications of a GSC-like structure with postfiltering [6, 7, 8] is developed. Thus, these algorithms obtain a good performance against both directional and diffuse noises.

In figure 2, a diagram of the new proposed integrated adaptive Wiener-filtering structure is shown. It can be seen that a Time Delay Compensation (TDC) block has been added to simulate target signal coming always from broadside.

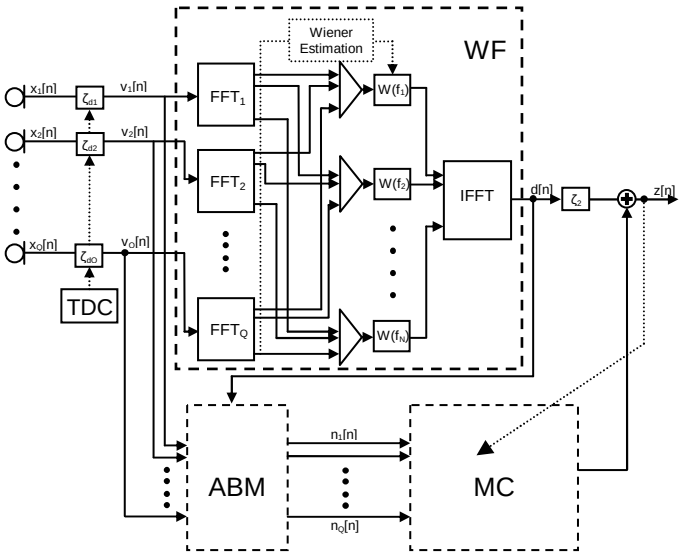

Figure 2. The IAW beamformer.

The newness of our proposal is that we integrate the stage of robust adaptive beamforming with ABM and the stage of postfiltering in a single whole. The way in which our structure does it consists in applying Wiener filtering to generate the output of the FBF. This filtered output is used by the ABM and by the MC. Therefore, the underlying idea behind the entire work is that the better the output of the FBF is, the better the ABM is estimated and moreover the MC will have better estimations of the reference target signal available and that of the reference noises at the output of the ABM.

Furthermore, it is expected that the integration of the Wiener filter inside a GSC-like structure may provide some other benefits. Usually, when noise is very high, postfiltering techniques can reduce it very well but in exchange for a more degraded signal. Our structure can reduce this degradation in two ways. First of all, the Wiener filter is directly applied to the FBF output, that is the signal for which the optimal Wiener is computed, and not to the output of the complete beamformer. And second, the MC can compensate some of the artifacts or degradations introduced in the signal by the filtering process.

5. Experiments and results

Three sets of different experiments have been carried out. In the first one we evaluate the noise and interference reduction capability in presence of an interfering speaker and also the speech degradation when only the target speaker is present. The second experiment presents speech recognition results using clean trained models when only a speaker is present. In both experiments array data is obtained convolving real microphone array impulse responses from the RWCP database [9] with clean signals and adding real microphone array ambient noise from the same database. The impulse responses and the noise used were recorded in a high reverberating meeting room (approximately 780 milliseconds of reverberation time) with a seven element microphone linear array of 5,66 cm of separation between microphones. Finally,

in the third experiment some preliminary results of distortion with real array data are shown. In all three cases, the proposed integrated adaptive Wiener-filtering (IAW) beamformer is compared to the Delay&Sum (DS), the DS postfiltered (DSP), the GSC with CCAF-LAF structure (GSC) and the same adaptive beamformer postfiltered (GSCP).

5.1. Speech enhancement results

In these experiments target speaker (male) is situated in the broadside and an interfering speaker (female) is at about 40 degrees. Speech used to simulate the array data are close-talking recordings of Spanish sentences of about 5 seconds length. Different signal to noise ratios (SNR) and signal to interfering ratios (SIR) were simulated and each beamforming technique has been simultaneously applied to the simulated 'complete' signal and also to the target signal alone, to a noise-alone signal and to an interference-alone signal in order to easily compute SNR and SIR gains.

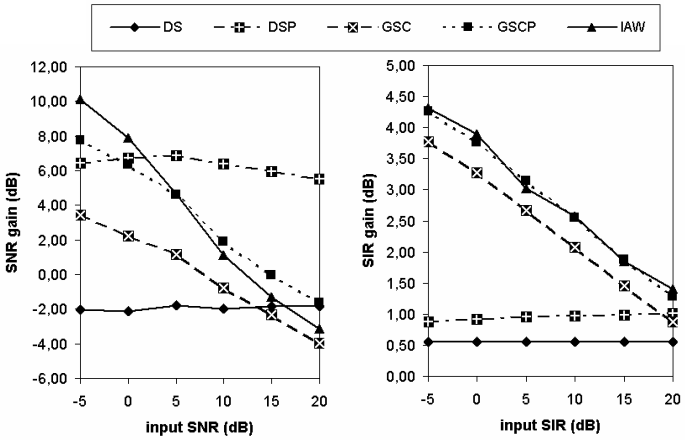

Figure 3. On the left side SNR gain and on the right side SIR gain.

Figure 3 shows on the left side SNR gain obtained for each technique for different input SNRs and with input SIR fixed to 15 dBs. On the right side, SIR gain for different SIRs at an SNR of 15 dBs is shown.

Regarding to the SNR gain results we can first notice that all the Wiener filtering based techniques (DSP, GSCP and IAW) usually present a better performance than the others, especially DSP beamformer. Furthermore, all GSC based techniques (GSC, GSCP and IAW) present a strong dependency with the input SNR and performs clearly worse as long as SNR increases. However, a great performance in very low-noise conditions is obtained with GSCP and IAW beamformers, even better than DSP. In that case, the proposed beamformer is especially good thanks to a better estimation of the noise references used by the multiple-input canceller.

On the SIR gain plot we can observe, as it could be expected, that only GSC-based beamformers show a considerable interference reduction performance. Concretely, GSCP and IAW are the best ones and present a similar gain. Moreover, a similar behaviour to the SNR gain is observed and interference reduction is more important when SIR decreases.

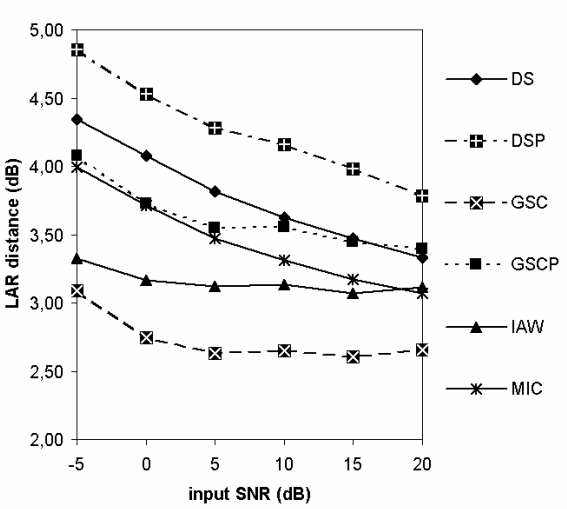

Figure 4. LAR distances for each beamformer and different input SNR when only the target speaker is present.

Log area ratio (LAR) distance, which is a well-correlated measure with subjective quality, is computed for the output of the beamformers and also for the fourth microphone arriving signal (MIC) for different SNRs in absence of the interfering speaker. Attending to results shown on figure 4 we can see that the proposed IAW beamformer clearly obtains better results than GSCP and DSP beamformers, but it is outperformed by GSC. This is not surprising as long as IAW beamformer is a Wiener filtering based technique and a greater degradation of the signal could be expected in exchange of a best performance against ambient noise and interference.

As preliminary conclusions, we can state that both GSCP and IAW beamformers show similar performance attending to noise and interference reduction, but IAW outperforms GSCP beamformer in very high noise conditions and presents lower LAR distance results. Therefore, IAW can be considered as a convenient beamformer for a wide range of conditions, particularly when noise is not very low.

5.2. Speech recognition tests

Speech recognition tests consist on the usage of the microphone array processing techniques under study as a front-end for a continuous digit recognizer. Recognition system was implemented with HTKv3.0 toolkit. A 39 feature vector was used composed by the static cepstrum representation and the delta and acceleration parameters. CDHMM of digits with 18 states and 3 mixtures for each state was trained with the clean speech training set of the Aurora1 database. Array data for testing was obtained from part of the Aurora1 test set convolving it with the impulse responses and adding the real ambient noise at different SNRs. In table 1, columns with accuracy results from 20 to 0 dBs of SNR and a column with the average results (AVER) are shown for all the tested beamformers and also for the fourth microphone of the array (MIC).

	20	15	10	5	0	AVER
MIC	58	55,66	49,55	37,27	17,96	43,69
DS	63,56	60,95	54,53	43,14	24,04	49,24
DSP	58,21	57,02	53,76	47,53	36,72	50,65
GSC	67,98	65,31	60,12	48,39	28,74	54,11
GSCP	60,91	58,7	54,28	47,25	35,52	51,33
IAW	62,82	60,98	57,75	52,32	43,2	55,41

Table 1. Accuracy (%) speech recognition results.

Although the proposed beamformer is the one that obtains the highest average results, only a marginal improvement is obtained for speech recognition. In fact, in high SNR situations the proposed technique performs clearly worse than others, but this bad performance is extensive to all Wiener filtering based techniques. Hence, it seems that distortion introduced affects stronger the performance of the speech recognition system than the influence of the noise reduction obtained when SNR is not low enough. In fact, these speech recognition results are well-correlated with the observations of previous section where IAW was also the best beamformer in very high noise conditions but it was outperformed in low noise presence.

5.3. Experiments with real array data

Experiments with real array data must be done to verify usefulness of the new approach for both speech enhancement and speech recognition. In this section some preliminary experiments with real microphone array recordings from the CMU array database [10] are reported. Seven microphones linearly distributed with an inter-sensor separation of 4 cm were used. Two experiments with different distances of 1 meter and 3 meters from the speaker to the array were done to evaluate speech enhancement obtained by the different methods.

Figure 5 shows LAR distances for each beamformer and also for the fourth microphone (MIC) of the array. Very good results are obtained by the proposed IAW beamformer and it is the best one among all the techniques under study. These results with real array data, besides some of the good results shown in previous sections, seems to corroborate our expectance that a better performance can be obtained if we integrate adaptive beamforming with Wiener filtering.

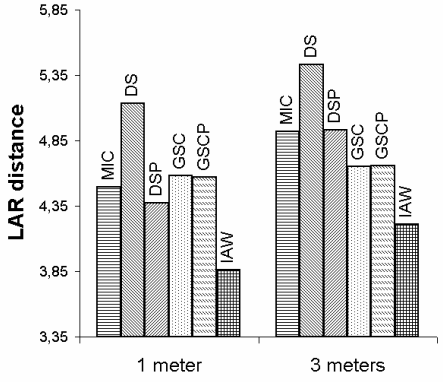

Figure 5. LAR distance results for a speaker located a 1 and 3 meters of the microphone array.

6. Conclusions

In this work we have presented a new robust adaptive beamformer with microphone arrays called integrated adaptive Wiener-filtering beamformer. Novelty of the proposal is that we integrate postfiltering techniques in the fixed beamformer path of a GSC-like structure beamformer with adaptive blocking matrix. In this way, a cleaner output of the fixed path is obtained resulting in a robust to noise and to interference scheme, besides a low level of degradation of the speech signal in a broad range of situations. Experimental results with simulated data in both speech enhancement and speech recognition tests at least confirm usefulness of this approach in most of the cases. Moreover, preliminary results with real data strengthen convenience of the proposed beamformer.

7. Acknowledgements

The authors would like to thank Professor Dr. Climent Nadeu and Dr. Jaume Padrell for their helpful advices. The work has been partially supported by the EUproject FAME (IST-2000-28323). Alberto Abad is supported by a Catalan Government research fellowship.

8. References

[1] Johnson D.H. and Dudgeon D.E., "Array signal processing", Prentice Hall, 1993.

[2] Brandstein M. and Ward D. (Eds.), "Microphone Arrays", Springer, January 2001.

[3] Griffiths L.J. and Jim C.W., "An alternative approach to linearly constrained adaptive beamforming", IEEE Trans. On Antennas and Propagation, vol. AP-30, pp. 27-34, January 1982.

[4] Marro C., Mahieux Y. and Simmer K.U., "Analysis of Noise Reduction and Dereverberation Techniques based on Microphone Arrays with Postfiltering", IEEE Trans. On Speech and Audio Processing, vol. 6, pp. 240-259, May 1998.

[5] Hoshuyama O. and Sugiyama A., "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix using Constrained Adaptive Filters", IEEE Proc. ICASSP'96, pp. 925-928, May 1996.

[6] Bitzer J., Simmer K.U. and Kammeyer K.D., "Multi-Microphone noise reduction by post-filter and super-directive beamformer", Proc. IWAENC'99, pp. 100-103, Sept 1999.

[7] McCowan I.A., Moore D. And Sridharan S., "Speech Enhancement using near-field superdirectivity with an adaptive sidelobe canceller and post-filter", Proc. 2000 Australian International Conference on Speech Science and Technology, pp. 268-273, December 2000.

[8] Fischer S. and Simmer K.U., "An Adaptive Microphone Array for Hands-Free Communication", Proc. IWAENC'95, pp. 44-47, June 1995.

[9] RWCP Sound Scene Database in Real Acoustic Environment. http://tosa.mri.co.jp/sounddb/indexe.htm

[10] Sullivan T. "Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition", Ph.D. thesis, August 1996. http://fife.speech.cs.cmu.edu/databases/micarray/