Test of Adaptive Beamformers for Speech Acquisition in Cars

Sofiène Affes Yves Grenier

Ecole Nationale Supérieure des Télécommunications
Télécom Paris, Département Signal, 46 rue Barrault, 75634 Paris Cedex 13, France.
E–mail: affes@sig.enst.fr, Tel: 33 1 45817782, Fax: 33 1 45887935.

ABSTRACT

This paper is devoted to a comparison of some of the representative techniques published in the literature of adaptive beamforming in real applications, when applied to speech acquisition in car acoustic environment. The evaluation results show unfortunately the weakness of these methods to reduce the noise in the tested environment. An acoustic characterization of the problem is hence made to explain this lack of performance. It reveals that the main limitation comes from the diffuse noise field. On the other hand, reverberation is shown to degrade speech intelligibility. Beamformers efficiently remove this effect.

1. Introduction

Speech acquisition in car environments is a difficult problem that receives growing interest, mostly because of hands-free mobile radio-communications [1]. For transmission, as well as for automatic speech recognition, it is necessary to properly pick up the speech signal and efficiently reduce the noise that corrupts it.

In this paper, we investigate the use of adaptive beamformers and microphone arrays for noise reduction and speech acquisition in cars. The knowledge of the exact or the approximate desired source position to be steered is a basic assumption in beamforming [1-3]. The undesired noise signals are hence known to be outside that look direction of interest. This is an underlying concept which motivates the choice of beamforming techniques for acoustic acquisition in hands-free telephony. The idea is to extract the desired source impinging from the driver's position, and to simultaneously reduce the rest of all possible undesired noise sources (engine noise, loudspeaker, aerodynamic noise, etc...).

Beamforming algorithms were however developed in radar and sonar processings. They can be used for speech processing, although there are some differences to take into account. The speech is indeed wideband and nonstationary. The spectral characteristics of the noise and speech are often identical. The distance between the source and the sensors is small, and the mobility of the speaker is not negligible. Reverberations and the presence of both spatially diffuse and point jammers represent an essential cause of desired signal distortion or cancellation, and unsatisfactory noise reduction [4]. Hence, modifications must be brought to classical beamformers to give more robust algorithms.

We have actually selected five beamformers and evaluated them under our test bench. Due to the physical complexity of car acoustic environment, special care has been taken in their selection for tests and evaluation. The tested algorithms are those developed by Kaneda and Ohga [5], by Sondhi and Elko [6], the post-filtering technique derived by Zelinski [7] and its modified version proposed by Simmer and Wasiljeff [8], and finally the multi-channel spectral subtraction developed by Gierl [9]. This leaves out several techniques, but most of them are very close to one of the selected algorithms briefly described in section 2.

Our evaluation is based upon two kinds of measurements: the improvement in SNR (signal to noise ratio), and the degradation of the extracted speech due to linear or nonlinear distortion. The tests were run on speech sentences uttered by several speakers, both male and female, for various realistic SNR's. All the results are presented in section 3.

A comparison with mono-sensor spectral subtraction techniques showed however that the gain provided by the microphone array is too weak to justify

[1] This study was partly funded by the EEC, under the European contract ESPRIT project 6166 FREETEL "Enhancement of Hands-Free Telephony". The corresponding part of the project ended in July 1993.

the increased cost. Therefore, an acoustic characterization of the noise in car and the propagation conditions is made in section 4 to analyze these unsatisfactory results.

This characterization shows the limitation of the filtering performed by beamforming regarding the coherence and both the spatial and frequency locations of the noise and speech. It shows however that beamforming is able to significantly increase the clarity index which specifies the quality of the acoustic paths between the speaker and the received signals. This allows a better intelligibility of speech at the listening.

Some other conclusions and recommendations are finally made in section 5.

2. Presentation of the beamformers

At time t, a reasonable model for speech recordings with an array of m microphones in adverse environment is given by:

$$x_{i,t} \quad = \quad g_{i,t} \otimes s_t + n_{i,t} \quad , \tag{1}$$

where $x_{i,t}$ and $n_{i,t}$ are respectively the noisy speech and noise signals received at the i^{th} microphone. s_t is the source signal issued from the speaker's mouth. $g_{i,t}$ is the impulse response between the source and the i^{th} microphone. The symbol $\otimes$ denotes here the linear convolution.

This model assumes that the disturbing noise is additive and does not tackle the problem of Lombard effect, where the speech signal is distorted due to the speaker's stress.

Now the aim is to recover the speech signal s from the filtering of the observations x_i by the m tapped delay-lines of the beamformer say h_i, where $i = 1, 2..., m$. For such a scheme, a typical constraint is the one that forces the array to have a pure delay response in the desired look direction:

$$\| R(f) - e^{-j2\pi\tau f} \|^2 = 0 \quad , \tag{2}$$

where:

$$R(f) \triangleq \frac{1}{m} \sum_{i=1}^{m} H_i(f) \times G_i(f) \quad , \tag{3}$$

is the total desired frequency response, and τ is the required time delay. $H_i(f)$ and $G_i(f)$ denote respectively the Fourier transforms of h_i and g_i. This constraint referred to as the distortionless constraint is however too rigid, and the improvement in SNR is generally not satisfactory.

The idea of Kaneda and Ohga is then to allow a small degradation of the speech signal, that remains subjectively acceptable to the human auditory system:

$$\| R(f) - e^{-j2\pi\tau f} \|^2 \leq \epsilon \quad , \tag{4}$$

where ϵ is a fixed positive threshold. Hence, it is proved in [5] that the new constraint improves noise reduction.

Noticing the relative insensitivity of speech quality to phase distortion, Sondhi and Elko introduced a softer constraint:

$$\left| \| R(f) \|^2 - 1 \right|^2 \leq \epsilon \quad . \tag{5}$$

$R(f)$ is hence approximated by an all-pass filter, and the noise reduction is shown to be more efficient [6].

This method assumes however that the received signals are already steered. This is actually equivalent to the following additional constraints:

$$\| Z_i(f) - e^{-j2\pi\tau f} \|^2 = 0 \quad , \quad (i = 1, 2, ..., m) \tag{6}$$

where:

$$Z_i(f) \triangleq U_i(f) \times G_i(f) \quad , \tag{7}$$

is the desired frequency response at the i^{th} microphone. $U_i(f)$ is the Fourier transform of u_i, the i^{th} response of the steerer. Then $R(f)$ is to be taken as the mean of the tapped-delay line responses $H_i(f) \times Z_i(f)$ for $i = 1, 2, ..., m$. These limiting constraints are actually assumed by all the following methods as well.

Assuming now the noise to be uncorrelated and diffuse, Zelinski proposed a simple Delay-Sum (DS) beamformer (*i.e.* $H_i(f) = 1$), which is known to be optimal for spatially diffuse noise reduction (for $m = 8$, $10\ log(m) \simeq 9\ dB$). The output of the DS beamformer is then processed by an adaptive Wiener post-filter to remove the residual noise still present as follows [7]:

$$W_Z(f) = \frac{2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{Re}\{Y_i(f) Y_j^*(f)\}}{m(m-1) \left\{ \frac{1}{m} \sum_{i=1}^{m} \| Y_i(f) \|^2 \right\}} \quad , \tag{8}$$

where for $i = 1, 2, ..., m$:

$$Y_i(f) \triangleq U_i(f) \times X_i(f) \quad . \tag{9}$$

Simmer and Wasiljeff observed however an overestimation factor of the power spectral density of the

noise in (8), and proposed a modified version of the post-processing as follows [8]:

$$W_{SW}(f) = \frac{2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{Re}\{Y_i(f) Y_j^*(f)\}}{m(m-1) \, \| \frac{1}{m} \sum_{i=1}^{m} Y_i(f) \|^2} \quad . \quad (10)$$

Finally, Gierl proposed a multi-dimensional spectral subtraction method implemented in a GSC-like structure [9]. $\frac{m(m-1)}{2}$ difference channels are then formed to adaptively estimate and remove the noise present in the corresponding sum channels.

3. Evaluation Results

We have recorded a database of speech signals and noise in a car environment using the microphone array described in [10]. This database contains 3 noise files recorded at various speeds of 60 km/h, 90 km/h and 130 km/h. It also contains 16 clean speech files of both male and female speakers, recorded with the car stopped. The clean signals allow by addition with the noise alone to build artificial experiments that remain realistic, and for which we can control the SNR. The clean speech files are hence corrupted by the noise files at the SNR's of 10 dB, 3 dB and 0 dB corresponding to the increasing speeds.

We hence compute the gain in SNR in dB and the distortion of the desired source signal in %. The mean and covariance over the 16 files are then given for each method at the different speeds.

Speed	$E(G_{dB})$	$\sigma(G_{dB})$	$E(D_{\%})$	$\sigma(D_{\%})$
60 km/h	-6.7	2.0	73.9	26.0
90 km/h	-1.1	1.8	60.1	12.4
130 km/h	2.4	2.0	68.3	12.7

Table 1. Mean and standard deviation of SNR gain and distortion of AMNOR for various speeds.

Speed	$E(G_{dB})$	$\sigma(G_{dB})$	$E(D_{\%})$	$\sigma(D_{\%})$
60 km/h	-1.7	1.7	45.0	18.1
90 km/h	1.5	1.6	31.1	7.1
130 km/h	5.3	1.1	22.0	3.8

Table 2. Mean and standard deviation of SNR gain and distortion of the method of Sondhi and Elko for various speeds.

The objective results given by AMNOR show that the source signals underwent severe distortion, and that the noise reduction is unefficient (see table 1). In comparison to this method, the algorithm of Sondhi and Elko effectively allows some phase distortion of the source signals, with however better capabilities of noise reduction and less distortion (see table 2). We notice in both methods that the gain G increases at a higher speed, while the distortion D decreases. This is precisely the expected effect of the LMS-based beamformers which better reduce the noise with less distortion at lower SNR's [4].

The results remain however unsatisfactory. First, this can be explained by the fact that the beamformers have been trained or used in a simple LMS scheme with nonstationary correlated noises. These noises cause the LMS algorithm to have very slow modes of convergence, which require training periods longer than those actually used. This point can be however partly overcome by the reduction of the filter size, and the use of improved versions of LMS. Second, these algorithms assume the presence of spatially located jammers to justify the choice of adaptive beamforming. We will see in the next section that the noise is rather diffuse. This makes both algorithms far from their optimal performances.

On the other hand, the 3 last algorithms assume the noise to be spatially diffuse. This assumption seems to be more adequate to the studied noisy environment, and likely allows better results.

Speed	$E(G_{dB})$	$\sigma(G_{dB})$	$E(D_{\%})$	$\sigma(D_{\%})$
60 km/h	7.9	1.9	10.8	2.8
90 km/h	6.5	0.9	11.3	2.6
130 km/h	6.4	0.4	12.0	2.5

Table 3. Mean and standard deviation of SNR gain and distortion of the method of Zelinski for various speeds.

Speed	$E(G_{dB})$	$\sigma(G_{dB})$	$E(D_{\%})$	$\sigma(D_{\%})$
60 km/h	5.1	1.3	4.9	0.8
90 km/h	4.6	0.7	7.0	1.2
130 km/h	5.1	0.3	8.3	1.7

Table 4. Mean and standard deviation of SNR gain and distortion of the method of Simmer and Wasiljeff for various speeds.

Indeed, the post-filtering techniques achieve a higher gain in SNR with relatively small amount of distortion. The method of Zelinski particularly reduces the noise better than the method of Simmer and Wasiljeff, with however more distortion (see tables 3 and 4). These results actually illustrate the effect of the noise overestimation factor resulting from the Wiener post-filtering proposed by Zelinski. The modified version of this filter proposed by Simmer and Wasiljeff precisely avoids this effect basically at high SNR's. We notice for this method that the gain G is quite stable at various speeds, while it tends to decrease and to saturate at low SNR's with a comparable level for the method of Zelinski.

In comparison to the post-filtering techniques, Gierl's method yet achieves a higher gain in SNR with

Speed	$E(G_{dB})$	$\sigma(G_{dB})$	$E(D_{\%})$	$\sigma(D_{\%})$
60 km/h	7.3	2.7	14.9	3.0
90 km/h	8.4	1.9	23.1	3.9
130 km/h	10.8	1.6	29.3	4.0

Table 5. Mean and standard deviation of SNR gain and distortion of the method of Gierl for various speeds.

more distortion. This is of course the expected and immediate effect of the spectral subtraction. This technique is well characterized by the presence of musical tones at the listening. This artifact is very unpleasant as much as the noise is nonstationary, and significantly limits the interest in this method usually motivated by a higher gain in SNR.

All these results are confirmed by subjective tests. The processed files were actually ranked between 1 to 5 by 25 listeners, naive and experts. The ranking given by the objective and subjective tests showed a high coherence. The results are clearly in favor of the post-filtering techniques: the algorithm proposed by Simmer and Wasiljeff closely followed by the method of Zelinski.

However, the global results obtained in these simulations are perhaps disappointing since they are not extremely high in terms of noise reduction. Nonlinear methods like spectral subtraction achieve better SNR's. Rejecting beamforming techniques from this conclusion would not be satisfactory.

In fact beamforming techniques produce other effects than mere noise reduction. A method like Zelinski's or Simmer and Wasiljeff's methods reduces the noise with a very small distortion. Such a treatment can be followed by spectral subtraction that will work in a better way, that is, for lower initial SNR's. The beamformer would then increase the level of noise that may be accepted by the spectral subtraction technique.

4. Acoustic Characterization

The evaluation of the beamformers has led us to the results described in section 3. The performance of various beamformers was not satisfactory. We tried to exhibit the physical origin of the limitations that the beamformers faced. The acoustic environment in the car is characterized by two aspects: the noise level is very high, and the propagation is complicated by the proximity between the source and the sensors, and by the size of the volume inside the car that has dimensions comparable to the wavelengths. A part of our work was then devoted to the characterization of the acoustic environment.

The acoustic characteristics that are relevant to our problem are: the acoustic path between source and sensor, the frequency content of the noise and the spatial distribution of the noise. To access to acoustic paths, we measured the impulse responses between a loudspeaker located at the speaker position and the sensors. The measurement was done by sending Golay codes through the loudspeaker, and convolving the signals picked by the sensors with code sequences. The noise was characterized by the spectral estimation of the signal on one sensor. We use parametric techniques (smoothed or averaged periodograms) as well as parametric techniques (maximum entropy or minimum variance).

The characterization of the spatial distribution was done in two steps. First, the power spectral density matrix of the set of sensor signals was estimated using smoothed periodograms. Second, at each frequency, spatial analysis was performed on the power spectral density (or covariance) matrix using narrowband source location techniques. Other tests were also applied to the spatial covariance (*e.g.* sphericity tests) in order to detect spatially diffuse noise.

The acoustic characterization shows that the noise inside the car is spatially incoherent at low frequency (below 800 Hz). This is precisely the frequency region where the energy of the noise is high. On the opposite, the noise at high frequency tends to be more localized (no point sources, but limited regions that radiate) at high frequency. Then the efficiency of the beamformer starts to be acceptable above 800 Hz, but the interest is limited by the fact that the noise at these frequencies is less noticeable.

Beamformer on the other hand may have some interest. Indeed, we have also characterized the propagation of sound inside the car, by measuring the impulse responses of a large set of acoustic paths, from one point to another inside the car. One conclusion is that the clarity index associated to each path, which specifies the quality of an acoustic channel for speech transmission is too low. It is the ratio of the total energy of the impulse response to the energy contained in the late reverberation part of this response say h:

$$C(h) \triangleq 10 \log_{10} \left(\frac{\sum_{t=0}^{\infty} h^2(t)}{\sum_{t=T_d}^{\infty} h^2(t)} \right), \qquad (11)$$

where T_d is the total duration of the direct path and the early reflections (see figure 1-c). A consequence is that the speech picked by microphones that are far from the speaker mouth (e.g. a microphone located on the dashboard) will not sound pleasant to the listener. We have shown that this was greatly improved by the beamformers.

Indeed, the beamformer is able to increase the clarity index. The quality of speech transmitted is con-

sidered as good when this clarity index exceeds 10 to 12 dB. The measurements done in [11] show that hands-free systems never reach this level when using a single microphone. We shall show in this section that beamforming techniques increase the clarity index to a level as high as 20 dB.

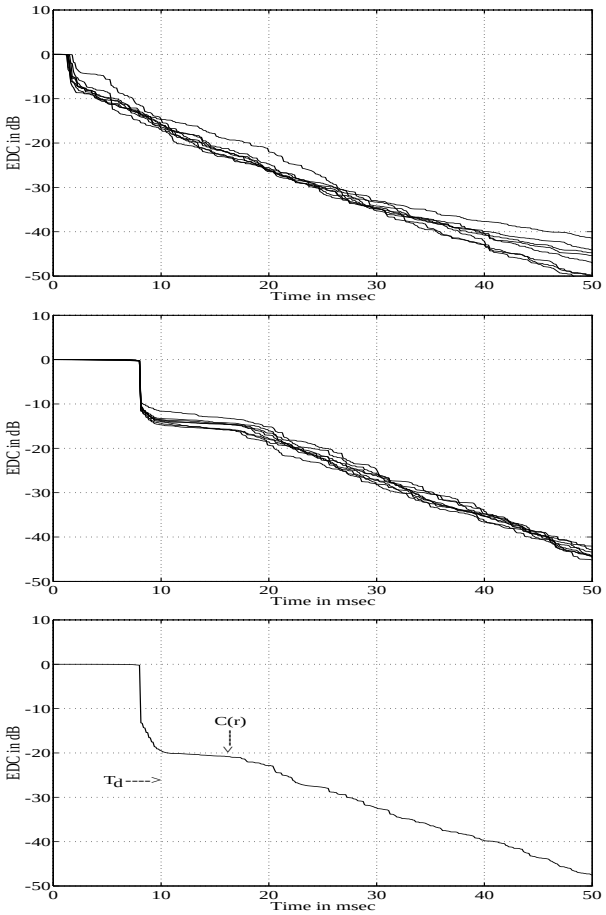

Figure 1. Normalized Energy Decay Curves. top: impulse responses - middle: steered responses - bottom: total response at sum.

The part of the beamformer that is most concerned by this effect is the synchronization of the signals. This is in fact a first deconvolution step that reduces the effect of the channel. The efficiency of the synchronization of the received signals by the proposed steering unit is shown by figures 1-a,b,c where we plot the Energy Decay Curve defined as follows for an impulse response say h:

$$E_h(t) \triangleq \sum_{\tau=t}^{\infty} h^2(\tau). \qquad (12)$$

In figure 1-a, we can see that the channel impulse responses have a clarity index ranging from 4 to 8 dB, with relatively strong reflections and reverberations. The figure 1-b shows that the synchronized impulse responses using the proposed steering unit have a significantly enhanced range of 12 to 15 dB for the clarity index. Reflections and reverberations are noticeably reduced, with perfect synchronization of the responses to an equivalent delay of 8.125 msec. The filter resulting from the sum of the synchronized responses prior to any further processing (Delay-Sum beamforming) has a clarity of 20 dB, with an efficient total reduction of reflections and reverberations (see figure 1-c).

5. Conclusion

In this paper, we have evaluated the ability of adaptive beamformers to reduce the noise in speech signals for hands-free radio-mobile telephony.

The evaluation results and the acoustic characterization prove that noise reduction is a difficult task for beamformers in the studied environment. Since the noise turns out to be spatially diffuse, a simple DS beamformer is likely to achieve better results.

They show however their capacity or potential to steer or synchronize the source signals. Subjective tests particularly prove that the listener is more sensitive to signal distortion than to noise reduction. These final statements underline the importance of the steering phase which is precisely not emphasized by most of the authors. Moreover, a post-filtering step like the one proposed by Simmer and Wasiljeff becomes necessary as the performance of DS is bound to 9 dB for spatially diffuse noises.

However, the evaluation results do not reflect the real efficiency of a "good" steering: we only identified the inverse filters of the impulse responses we have measured. It is not readily defined that the resulting steerers really achieve the required quality of synchronization: the measured responses can be corrupted or distorted. In addition, the motion of the speaker is not taken into account. Hence, an adaptive procedure should be proposed for the steering of the signals.

Simmer *et al.* [13] already proposed a steering unit based on the correlation, which is not likely to work properly in the studied environment. We alternatively worked on a wideband robust adaptive beamforming algorithm allowing a continuous time delay steering of the speech sources [14]. At present, we are trying to better adapt this algorithm to the studied acoustic environment.

References

[1] B. Widrow *et al*, "Adaptive noise cancelling: principles and applications", *Proc. of IEEE*, Vol. 63, No 12, pp 1692-1715, December 1975.

[2] W.F. Gabriel, "Adaptive arrays - an introduction", *Proc. of IEEE*, Vol. 64, No 2, pp 239-272, September 1976.

[3] B.D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering", *IEEE ASSP Magazine*, pp 4-24, April 1988.

[4] B. Widrow *et al*, "Signal cancellation phenomena in adaptive antennas: causes and cures", *IEEE Trans. on ASSP*, Vol. 30, No 3, pp 469-478, May 1982.

[5] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction", *IEEE Trans. on ASSP*, Vol. 34, No 6, pp 1391-1400, December 1986.

[6] M.M. Sondhi and G.W. Elko, "Adaptive optimization of microphone arrays under a nonlinear constraint", *ICASSP-86*, pp 981-984, Tokyo, Japan, April 7-11, 1986.

[7] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms", *ICASSP-88*, pp 2578-2581, New York City, USA, April 11-14, 1988.

[8] K.U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain", 2^{nd} *Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux-Technopolis, France, September 30 - October 2, 1992.

[9] S. Gierl, "Noise reduction for speech input systems using an adaptive microphone-array", 22^{nd} *ISATA*, pp 517-524, Florence, Italy, May 1990.

[10] Y. Grenier, "A microphone array for car environments", *ICASSP-92*, Vol. 1, pp 305-308, San Francisco, USA, March 23-26, 1992.

[11] Y. Grenier, Ed., *Characterization of the environments*, Deliverable 2.2, ESPRIT Project 6166 FREETEL, ENST-ARECOM, Paris, France, July 1993.

[12] P.A. Naylor and O. Tanrikulu, Ed., *Basic algorithms for noise reduction*, Deliverable 4.123.1, ESPRIT Project 6166 FREETEL, Imperial College, London, UK, July 1993.

[13] K.U. Simmer, P. Kuczynski and A. Wasiljeff, "Time Delay Compensation for Adaptive Multichannel Speech Enhancement Systems", *ISSSE'92*, pp 660-663, Paris, France, September 1-4, 1992.

[14] S. Affes, S. Gazor and Y. Grenier,"Wideband robust adaptive beamforming via target tracking", 7^{th} *SP Workshop on SSAP*, pp 141-145, Québec City, Canada, June 26-29, 1994.