Audio Engineering Society

Convention Paper 6329

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

DSP-IMPLEMENTED BROADBAND SUPERDIRECTIVE MICROPHONE ARRAY WITH AUDIBLE NOISE SUPPRESSION

Jose-Luis Sanchez-Bote[1], Joaquin Gonzalez-Rodriguez[1], and Javier Ortega-Garcia[1]

[1] Speech and Signal Processing Group (ATVS), http://www.atvs.diac.upm.es,
DIAC - E.U.I.T. Telecomunicación, Universidad Politécnica de Madrid, Madrid, 28031, Spain
jbote@diac.upm.es, jgonzalez@diac.upm.es, jortega@diac.upm.es

ABSTRACT

In this paper, a novel microphone linear array is proposed and implemented for real-time processing, working on a DSP processor in the frequency domain. The array, which is composed of 15 microphones in nested configuration, combines two multichannel techniques for speech improvement: SuperDirective beamforming (SD) and Audible Noise Suppression (ANS). The SD beamforming technique is an alternative to conventional or Delay and Sum beamforming (DS) with has worse low frequency spatial selectivity. ANS processing is based on the masking properties of the human auditory system and can benefit the perceived and objective quality of the processed signal. Although it has been successfully used in single channel systems, an enhanced multichannel version has been developed here, taking advantage of the extra information available in the acoustic spatial samples from the microphone array. Several on-line experiments are described here, assessing the real-time prototype.

1. INTRODUCTION

As a consequence of the availability of faster and cheaper digital signal processors, able to manage multichannel real-time processing, acoustic array signal processing can be considered a consolidated technique after its continuous development in the last 20 years [1]. When speech signal is transmitted to a far microphone, the recorded signal is degraded by two different components: noise and reverberation. Most approaches to enhance speech degraded by noise are based on spectral subtraction or Wiener filtering. Some works on single channel speech enhancement have tested the excellent performance of noise filtering using the masking properties of the human auditory system [2], causing the Audible Noise Suppression (ANS) method. The ANS technique here used is based on processing the noise components according its relative level, below or over the subjective audible thresholds, which are evaluated in each auditory critical band. The main contribution of this work is the use of a multichannel system, which after SD beamforming, where reverberation is severely reduced, can perform a better calculation of the masking thresholds from estimated noise-free speech signal, than can be closely obtained

by combination of the different spatial samples extracted from the enclosed sound field in the room. ANS method improves subjective perception of distortion and residual noise, consequence of short time speech processing. An important detail of the microphone array here presented is that it has been implemented on a DSP-based platform, allowing real-time processing of 15 synchronous audio channels, which includes frequency-domain SD beamforming and auditory-based postfiltering.

2. THEORETICAL BASIS AND SYSTEM DESIGN

2.1. SuperDirective (SD) beamforming

Beamforming techniques [1] are usually applied to extract and enhance speech signals in an acoustic space. Focusing the main lobe of the directivity pattern to a desired speech signal, noise and reverberation will be severely attenuated.

Consider a noise M-row vector $\mathbf{N}(\omega)$ picked up by a linear array of M microphones, and expressed in the frequency domain. The beamformed output $Y_{SD}(\omega)$ of the superdirective array, when were electronically aimed to a generic angle θ, is a linear combination of the M outputs:

$$Y_{SD}(\omega) = \mathbf{W}(\omega)^H \mathbf{A}(\omega) X_0(\omega) + \mathbf{W}(\omega)^H \mathbf{N}(\omega) \quad (1)$$

where $\mathbf{W}(\omega)$ is an M-row vector of beamforming filters, one per microphone, $\mathbf{A}(\omega)$ is the steering vector that represents all phenomena related with the acoustical-electrical conversion from the acoustic pressure to output signals, and $X_0(\omega)$ as the best achievable signal in a noise and reverberation free scene. Angle dependencies (θ) in $\mathbf{W}$ and $\mathbf{A}$ have been suppressed for the sake of simplicity.

With noise and reverberation, minimizing global power from non-principal directions is desired. The procedure, called here SD beamforming [1], that minimizes output power, attenuating noise and reverberation by enhancing low frequency directivity, for a distortion-less response in the main direction, is known as *Minimum Variance Distortionless Response* (MVDR) [1].

The distortionless condition guarantees a beamformed output $Y_{SD}(\omega)$:

$$Y_{SD}(\omega) = X_0(\omega) + N_{SD}(\omega) \quad (2)$$

where $N_{SD}(\omega)$ is the beamformed output of the noise. For the MVDR beamformer, it has been considered that the beamformed output is $Y_{SD}(\omega) = X_0(\omega)$ or reference, by application of the non-distortion condition when noise is absent. A known implementation [1] of a SD array, is represented in Figure 1. Here, the beamformed output of the SD array is given by:

$$Y_{SD}(\omega) = \underbrace{\mathbf{W_S}^T \mathbf{Y_D}(\omega)}_{Y_{DS}(\omega)} - \overbrace{\mathbf{H}^H(\omega)\underbrace{[\mathbf{B}\,\mathbf{Y_D}(\omega)]}_{\mathbf{Y_B}(\omega)}}^{Y_{BBF}(\omega)} \quad (3)$$

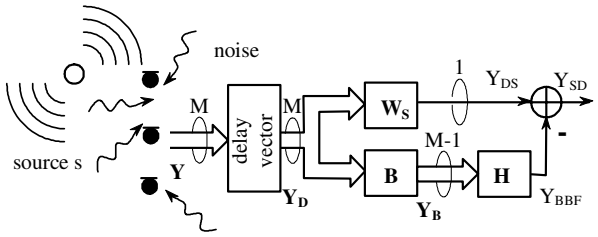

Figure 1 SD array using a blocking matrix.

The first addend $Y_{DS}(\omega)$ is the output of the conventional DS beamformer ($\mathbf{W_S}$ is a column vector with M coefficients of similar 1/M values), while the second addend gives the blocking output $Y_{BBF}(\omega)$, which includes information of noise and reverberation components not present in the main pointing direction. $\mathbf{B}$ is the blocking matrix that implements between-channels combinations in order to obtain minimum response toward the source for the elements of $\mathbf{Y_B}(\omega)$. A well-known blocking matrix is the Griffiths-Jim matrix [3] that combines M-1 acoustic dipoles (bidirectional pairs), choosing the M-1 consecutive microphone couples of the array. With the implementation described by (3) and represented in Figure 1, all time-aligned array channels ($\mathbf{Y_D}$) are available for post-processing. The choice of the multichannel filter $\mathbf{H}(\omega)$ as,

$$\mathbf{H}(\omega) = \left[\mathbf{B}\,\mathbf{\Gamma'}_{\mathbf{NN}}(\omega)\,\mathbf{B}^T\right]^{-1} \mathbf{B}\,\mathbf{\Gamma'}_{\mathbf{NN}}(\omega)\,\mathbf{W_S} \quad (4)$$

guarantees a MVDR response $-\mathbf{\Gamma'}_{\mathbf{NN}}(\omega)$ is the time-aligned noise coherence matrix–. Frequently a small scalar $\mu(\omega)$ is added [1] in the diagonal of $\mathbf{\Gamma'}_{\mathbf{NN}}(\omega)$ to avoid excessive non-coherent noise amplification:

$$\mathbf{\Gamma'}_{\mathbf{NN\mu}}(\omega) = \mathbf{\Gamma'}_{\mathbf{NN}}(\omega) + \mu(\omega)\,\mathbf{I} \tag{5}$$

This case is known as the restricted MVDR solution.

2.2. Audible Noise Suppression (ANS)

Noise reduction in single channel speech enhancement using postfiltering techniques has been largely studied, and different proposals exist [1]. Working with microphone arrays, the availability of multiple spatial samples of the acoustic sound field will help to obtain better estimates of clean speech signal, free of diffuse and coherent noise components as well of reverberation, resulting in an optimized postfiltering. If a postfilter with transfer function $H_{post}(\omega)$ is applied to the beamformed output $Y_{SD}(\omega)$, the clean speech can be estimated from:

$$\hat{X}_0(\omega) = H_{post}(\omega)\, Y_{SD}(\omega) \tag{6}$$

where, in case of magnitude-only processing, $H_{post}(\omega)$ is real. Many are the proposals for $H_{post}(\omega)$ from optimal Wiener filter to classical spectral subtractors [4].

When short time spectral analysis is used to implement the filter $H_{post}(\omega)$, a typical distortion known as "musical noise" is usually present, because of the differences between the noise averages available and the instantaneous spectrum to be subtracted or filtered. To increase subjective impression, ANS filtering has been successfully applied in single channel speech enhancement [2]. The ANS method is based on critical auditory band processing, where noise suppression is applied in agree with the threshold where the noise remains audible and called masking threshold $T(\omega)$. The masking threshold $T(\omega)$ is obtained for every auditory critical band, and shows the upper bound below where noise remains masked by signal. If the masking threshold is high, no much noise suppression is needed and vice versa, minimizing the musical noise and obtaining a drastic improvement in the perceived quality.

The ANS method has been extended by the authors [5] to work with multichannel audio data. A 2-stage multi-channel perceptual speech enhancement is proposed:

1. Computation of a clean signal estimate $Y_W(\omega)$ through a multichannel Wiener filtering [6] (see Figure 2).
2. ANS filtering of the beamformed signal $Y_{SD}(\omega)$ (Figure 2).

An intermediate step of speech enhancement is needed because $T(\omega)$ must be obtained from an estimate of the clean speech signal. A modified version $H_W(\omega)$ of multichannel Wiener filtering, proposed in [5]-[6], has been here used. The clean speech estimate $Y_W(\omega)$ at the output of the multichannel Wiener filter is now used to compute $T(\omega)$. The ANS filter, in the role of H_{post}, is given by:

$$H^2_{ANS}(\omega) = \frac{\Phi_{Y_{SD},Y_{SD}}(\omega)}{\alpha(\omega)\beta + \Phi_{Y_{SD},Y_{SD}}(\omega)} \tag{7}$$

where the parameter $\alpha(\omega)$ is obtained from the previously estimated masking thresholds $T(\omega)$:

$$\alpha(\omega) = \left[\hat{\Phi}_{N_{SD},N_{SD}}(\omega) + T(\omega)\right]\left(\frac{\hat{\Phi}_{N_{SD},N_{SD}}(\omega)}{T(\omega)}\right) \tag{8}$$

and where $\hat{\Phi}_{N_{SD},N_{SD}}(\omega)$ is the estimated noise autos-pectrum at the beamformer output, obtained in non-speech frames. For this task, the two-pole segmentation method [4], extremely efficient in single channel speech enhancement, has been used. The $\alpha(\omega)$ term is a measure of the noise to masking threshold ratio, and the $\beta \geq 1$ constant is an authors' proposal used to provide additional noise over suppression.

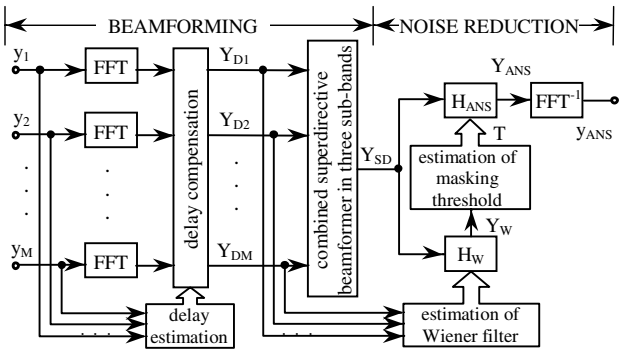

Figure 2 SD array using a blocking matrix.

3. PROTOTYPE IMPLEMENTATION

In this work, a real-time prototype of a 15-microphone nested linear SD microphone array [5] (Figure 3) with ANS processing has been implemented, tested and assessed.

The DSP board is the BWS PCI C6600 from Blue Wave

Systems, which includes one Texas Instruments TMS320C6701 floating point DSP (close to 1 GFLOP) and the multichannel input/output audio board BWS-PMC-16IO2, with 16 high quality (20 bits) analog inputs, and two analog outputs (24 bits).

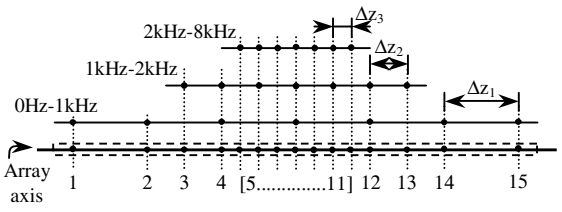

Figure 3 Three octave band nested array. $\Delta z_1 = 16$cm, $\Delta z_2 = 8$cm and $\Delta z_3 = 4$cm. The bands selected by each sub-array are expressed.

An additional reference lapel microphone, to record the close talking speech, has been added to the system as performance assessment. All 16 sensors are high quality AKG-C417 omnidirectional microphones. Additionally, a web cam has been placed over the array targeted to the front operational area. All reported applications allow mouse-based real time source tracking by a human operator from the web-cam video stream. In Figure 4 two images of the final prototype are depicted.

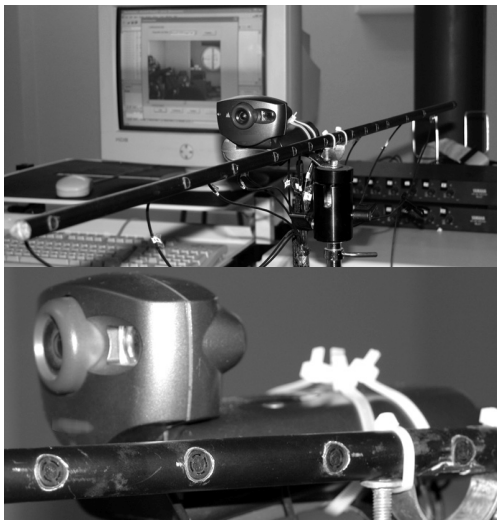

Figure 4 Final prototype. Back, the microphone amplifiers and the aiming screen on the PC monitor.

Input signals are 512pt.-Hanning windowed (32ms, $f_s = 16$kHz), with 67% overlapping (341 samples) for perfect reconstruction of time signal. After FFT computation, time delay and amplitude compensation is frequency-domain performed, where aligned channels are used for modified Wiener filter estimation, $H_W(\omega)$. The beamforming, spectrum is splitted into 3 sub-bands [0–1kHz], [1–2kHz] and [2–8kHz], where the first one is SD beamformed as in Figure 1 with $\mu = 0.0316$ and the two latter are DS beamformed. Finally, the output is ANS-postfiltered as described before in (7) and (8).

4. EXPERIMENTS AND RESULTS

4.1. Beamforming and directivity experiments

Beamforming abilities of the real time prototype (previous to ANS postfiltering) have been tested in an anechoic room. A pink noise sound source has been placed near (1.2m away) the array center, in order to minimize low frequency standing waves picking up. A Brüel&Kjäer turntable performs a 360º array revolution. In Figure 5, the directivity contour map D(θ, f), measured for the nested array proposed, is represented (results are shown just up to 180º because of symmetry). Here, the array is aimed to $\theta_0 = 0$º (endfire). The low frequency band has been beamformed with the SD method described before in (3), (4) and (5) and using $\mu = 0.0316$ (equivalent to -15dB).

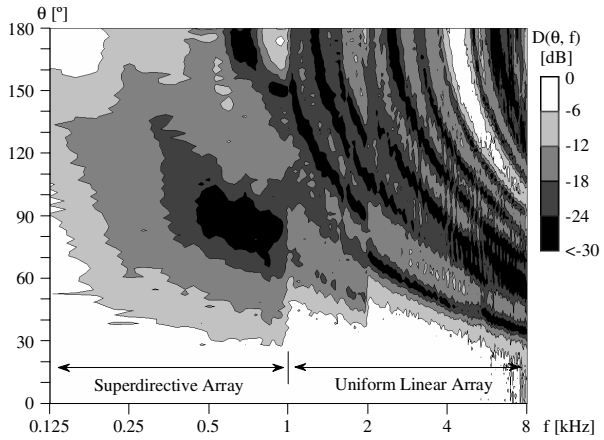

Figure 5 Directivity D(θ, f) contour map measured for the nested array proposed, aimed to $\theta_0 = 0$º (endfire) and with the source at $r_0 = 1.2$m. The low frequency band has been beamformed using the superdirective method as in (3)-(5), with $\mu = 0.0316$ (-15dB).

As depicted in Figure 6, the agreement from predictions to measures has been extraordinary, except for very low frequencies, where the combined influence of anechoic

room standing waves and loudspeaker directivity (experimentally reported in the over 150Hz), does not permit easy observation of the array performance. Different μ (5) values have been tested, concluding that the selected value $\mu = 0.0316$ (−15 dB) gives the better selectivity in the SD frequency band, [0–1kHz].

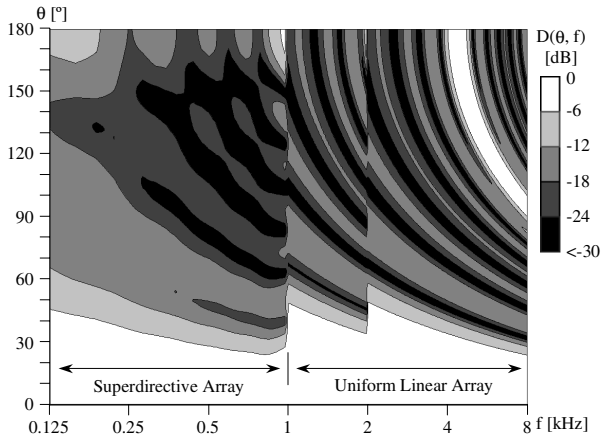

Figure 6 Theoretical directivity $D(\theta, f)$ contour map for the nested array proposed, aimed to $\theta_0 = 0^o$ (endfire) and with the source at $r_0 = 1.2$m. To be compared with Figure 5, where experimental measurements are shown.

4.2. Speech enhancement with ANS postfiltering

In order to test the real-time prototype, a highly reverberant room (Figure 7) has been selected ($T_{60} > 1$s in low frequencies). A real broadband Volvo car noise from the SpEAR Database (available at www.cslu.ogi.edu) has been selected and played back directed to the room walls.

Different experiments have been performed with different spatial configurations (endfire, as shown in Figure 7, and broadside aiming), different noise levels, and different suppression factors, β in (7). Results have been averaged over a speech database of 48 Spanish language utterances with three male and two female speakers.

Objective noise reduction tests have been performed, using the remaining noise $N_R(\omega)$ calculated as follows:

$$N_R(\omega) = Y_{ANS}(\omega) - X_0(\omega) \quad (9)$$

in similar way as (2) and with $X_0(\omega)$ known. $X_0(\omega)$ is the noise and reverberation free speech which is available. The Noise to Masking threshold Ratio NMR [2], gives a sign of audibility of the remaining noise, and the Gain in NMR, GNMR, is obtained as:

$$GNMR = NMR_{Y_8} - NMR_{Y_{ANS}} \quad (10)$$

where NMR_{Y_8} and $NMR_{Y_{ANS}}$ are respectively measured at the central (8th) microphone at the processor output. A sample of the experiments released is depicted in Figure 8 and Table 1.

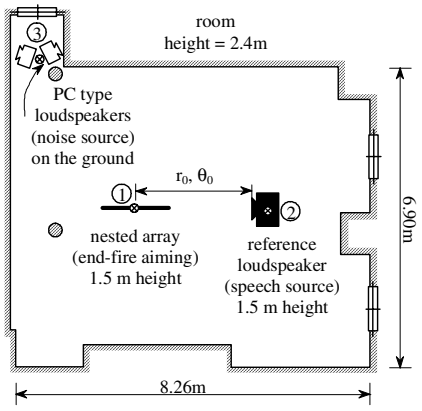

Figure 7 Room configuration for speech enhancement tests.

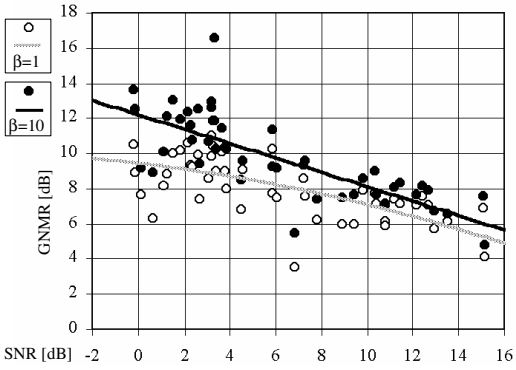

Figure 8 GNMR vs. SNR_{Y_8} for two values of over suppression factor β. Adjusted with 2nd-order polynomial.

In case of low signal quality (high noise and/or reverberation), higher β value are preferred which gives a clearer but less natural signal. When mid/good quality is available, $\beta = 1$ is better because the enhancement is similar to $\beta = 10$, but less distortion is present. It must be noted that SNR_{Y_8} in Figure 8 and Table 1 is the *a posteriori* SNR measurement from the central array cha-

nnel, obtained by considering average powers in speech activity frames, related with those ones in non-speech frames. This SNR calculation method is different to that performed in NMR of (10), where the remaining noise $N_R(\omega)$, necessary for calculation, is obtained from (9).

			GNMR[dB]		
β	θ_0 [°]	r_0 [m]	SNR<5dB	5dB<SNR<10dB	SNR>10dB
1	0	2	9.97	9.09	7.21
		4	8.91	5.68	6.36
	90	2	9.19	8.18	6.84
		4	8.94	6.55	6.03
10	0	2	12.47	9.95	7.97
		4	11.33	7.55	7.57
	90	2	11.17	9.29	7.36
		4	10.80	8.02	6.79
Partial result			10.29	7.82	6.85

Table 1 GNMR results for three SNR_{Y_8} steps.

5. CONCLUSIONS

A new multichannel auditory-based processor has been proposed, assessed and real-time implemented. The problem of reverberation and diffuse noise has been tackled through a SD beamformer in a sub-band configuration with a nested microphone array, obtaining high directivity in the full broadband of the array. Further noise reduction has been performed by a new proposal of a multichannel ANS processor, which takes into account the masking thresholds of the human auditory system in order to obtain a higher quality processed signal. A successful real-time implementation of the SD-ANS processor is now available for real-time testing or database acquisition in any real acoustic condition.

6. ACKNOWLEDGEMENTS

This work was supported with project TIC2000-1683-C03-02.

7. REFERENCES

[1] Brandstein, M. and Ward, D., *Microphone arrays*, Springer Verlag, Berlin, 2001.

[2] Tsoukalas, D.E., Mourjopoulos, J.N. and Kokkinakis, G., "Speech enhancement based on audible noise suppression", in *IEEE Trans. Speech Audio Processing*, vol.5, no.6, pp.497-514, 1997.

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

[4] Gay, S.L. and Benesty, J. (ed.), "Microphone Arrays", Chapter IV, in *Acoustic signal processing for telecommunication*, pp.181-282, Kluwer Academic Publishers, Massachusetts, 2000.

[5] Sanchez-Bote, J.L., Gonzalez-Rodriguez, J. and Ortega-Garcia, J., "A real-time auditory-based microphone array assessed with E-RASTI evaluation proposal", in *Proc. ICASSP 2003*, pp.481-84, Hong Kong, 2003.

[6] Gonzalez-Rodriguez J., Sanchez-Bote J.L. and Ortega-Garcia, J., "Speech dereverberation and noise reduction with a combined microphone array approach", in *Proc. ICASSP 2000*, pp.1037-40, Istanbul, 2000.