A REAL-TIME AUDITORY-BASED MICROPHONE ARRAY ASSESSED WITH E-RASTI EVALUATION PROPOSAL

J.L. Sánchez-Bote, J. González-Rodríguez, and J. Ortega-García
Speech and Signal Processing Group (ATVS)

DIAC - E.U.I.T. Telecomunicación - Universidad Politécnica de Madrid
Ctra. Valencia, km. 7 - Campus Sur, 28031 Madrid, SPAIN
jbote/jgonzalez/jortega@diac.upm.es - http://www.atvs.diac.upm.es

ABSTRACT(*)

In this paper, a real time nested microphone array based in the auditory properties of the human ear is presented. Three different stages in the development of the system are described. Firstly, the design of the new auditory-based microphone array is presented, obtaining better noise reduction using the masking properties of the human auditory system. Secondly, we show its validation through a new method called E-RASTI based in the well-known RASTI (Rapid STI – Speech Transmission Index -) intelligibility estimator. Additionally to classical enhancement *estimators as SNR, NMR (Noise to Masked Ratio) or AI* (Articulation Index), E-RASTI is proposed and validated for dereverberation assessment, used here with real speech signals and not with speech-like signals as in the original RASTI method. And finally as third stage, the real time implementation of this highly computing-demanding algorithm through the use of a DSP-based architecture based in the recent floating point TMS320C6701 is described.

1. INTRODUCTION

When speech signals are transmitted to a far microphone, the recorded signal is degraded by two different components, namely noise and reverberation. The different nature of these two perturbations makes necessary to tackle the problem with different processing alternatives. Most approaches to enhance speech degraded by noise are based on the assumption that signal and noise are fully uncorrelated. Filtering the short-time spectral amplitude of the degraded signal can efficiently eliminate or reduce the noise components, using spectral subtraction [1], or Wiener filtering [2][3]. However recent works in single channel speech enhancement have tested the excellent performance of noise filtering using the masking properties of the human auditory system [4][5][6]. This technique is known as *Audible Noise Suppression* (ANS) and is based on processing only those noise components that are over the subjective audible threshold, which are evaluated in each critical band. The ANS method improves subjective perception of the residual noise resulting from that processing, even though objective measurements of SNR show no significant improvements. The main contribution of this work is the use of a multichannel system that after beamforming, where reverberation is severely reduced, can perform better estimations of the masking thresholds of the noise-free speech signal from the different spatial samples extracted from the enclosed sound field in the room. This ANS multichannel system has been implemented in a DSP-based platform allowing real-time processing of the 15 synchronous audio channels sampled with 20 bits at 16 kHz, including frequency-domain beamforming and auditory-based postfiltering. Additionally, this paper introduces a new method called E-RASTI based in the well-known RASTI intelligibility estimator [9]. E-RASTI estimates speech quality by testing the modulation losses of speech signal intensity at very low frequencies with real speech signals instead of the speech-like signals used in classical RASTI, widely employed in speech intelligibility tests for room acoustics evaluation. Some dereverberation alternatives implemented in our laboratory have been tested in this paper through the use of E-RASTI, which has been shown to be an efficient estimator of the dereverberation abilities of the different array processors under evaluation.

2. MULTICHANNEL AUDITORY SPEECH ENHANCEMENT

In this paper we have used the masking thresholds of the human auditory system to control the amount of noise reduction that is necessary in each time frame. The underlying idea of the method is that noise only should be reduced down to the masking threshold of speech signal, which is related with the noise-free speech level in each critical band. If the noise level underpasses the masking threshold, computed as shown in [7], it may be considered as subjectively not audible. When the masking threshold is low, high noise suppression is needed, and vice versa. Lots of suppression functions using the masking threshold can be applied, having selected in our work that described in [5].

2.1. Noise-free speech estimation with microphone arrays

Masking thresholds should be obtained from clean speech, but in real systems the noise-free signal is unavailable and thresholds must be estimated. Wiener filtering with coherence modification [2], called Modified Wiener method (MW), has been used in this paper to get clean speech estimation.

The gain of this modified Wiener filter is:

(*) This work has been supported with project TIC2000-1683-C03-02

$$H_{MW}(\omega)=\begin{cases}\dfrac{\langle G_{xi\,xj}(\omega)\rangle-\langle G_{ni\,nj}(\omega)\rangle}{G_{xx}(\omega)} & if\, C(\omega)>CT \\ C(\omega)^{\alpha} & if\, C(\omega)<CT\end{cases} \quad (1)$$

$$\text{with} \quad C(\omega)=\frac{G_{x0}(\omega)}{\sqrt{G_{xx}(\omega)G_{00}(\omega)}} \quad (2)$$

where $C(\omega)$ is the interchannel coherence function, $\langle G_{xi\,xj}(\omega)\rangle$ and $\langle G_{ni\,nj}(\omega)\rangle$ are the averaged cross spectra over all channel pairs, the latter just considering time frames without speech activity. $G_{xx}(\omega)$ is the estimation of noisy speech autospectrum obtained by beamforming all channels in three frequency subbands, and α and CT (Coherence Threshold) are fixed parameters. In expression (2) $G_{x0}(\omega)$ and $G_{00}(\omega)$ are respectively the cross spectrum between the beamformed signal and the array central channel or reference channel and the autospectrum of the reference channel.

The main advantage of this configuration is that by using the multichannel information both coherent and non-coherent noise can be detected and processed differently. We use expression (1) to obtain a noise-free speech estimator as in the next equation:

$$\hat{Y}_{MW}(\omega)=H_{MW}(\omega)\cdot Y(\omega) \quad (3)$$

where $\hat{Y}_{MW}(\omega)$ is the estimation of the clean speech spectrum from Wiener filtering and $Y(\omega)$ is the noisy speech spectrum. $H_{MW}(\omega)$ can be used to obtain the final output (MW method) or may be used as clean speech estimator as follows.

2.2. Speech enhancement using masking thresholds

When a good estimation of a noise-free speech signal has been obtained, it is possible to apply the ANS method to reduce the noise to an inaudible level. So, the clean speech spectrum $\hat{Y}_{ANS}(\omega)$ can be obtained with:

$$\hat{Y}_{ANS}(\omega)=H_{ANS}(\omega)\cdot Y(\omega) \quad (4)$$

where $H_{ANS}(\omega)$ is the auditory enhancement filter from the ANS method. This filter is obtained as follows [5]:

$$H_{ANS}(\omega)=\frac{Y^{v}(\omega)}{a^{v}(\omega)+Y^{v}(\omega)} \quad (5)$$

where v is a parameter (normally fixed and not frequency dependent) and $a(\omega)$ is another parameter which is related with the masking threshold and then frequency dependent.

The masking threshold is represented by $T(\omega)$ and must be obtained from clean speech [estimated in $\hat{Y}_{MW}(\omega)$] as in [7]. The parameter $T(\omega)$ must modify $a(\omega)$ as follows: when the threshold is low compared to the noise, great attenuation should be done and therefore $a(\omega)$ must be high [see (5)]. Consequently, a comparison between the noise level and the speech level must be applied in every signal frame. Next expression verifies the latter:

$$a(\omega)=[N(\omega)+T(\omega)]\cdot\left(\frac{N(\omega)}{T(\omega)}\right)^{1/v} \quad (6)$$

with $N(\omega)$ as the noise autospectrum which can be estimated in non-speech frames. Although $T(\omega)$ has been written down as frequency dependent, it remains constant in each critical band, and $T(\omega)=T_b$ can be assumed, where b is the index associated with one critical band. The same assumptions are proposed for $a(\omega)=a_b$ and $N(\omega)=N_b$ (the noise is approximately constant in each critical band).

3. SYSTEM DESCRIPTION AND REAL-TIME IMPLEMENTATION

The microphone array arrangement used is shown in figure 1. As is shown, speech signal spectrum is split into three frequency subbands, each one from a microphone group.

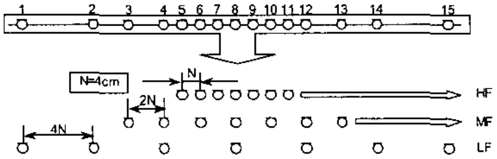

Figure 1: Nested array of fifteen microphones

In figure 2 the detailed processor scheme is presented. As can be seen, the multichannel signal may be processed in two ways, by ANS or MW methods (note that the MW output is used as clean speech estimate for the ANS system)

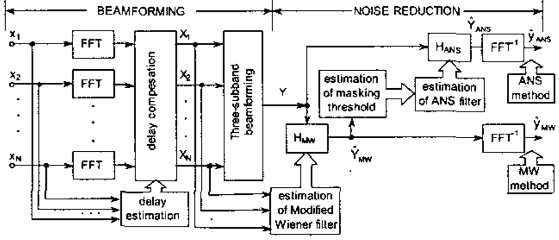

Figure 2: ANS method processor

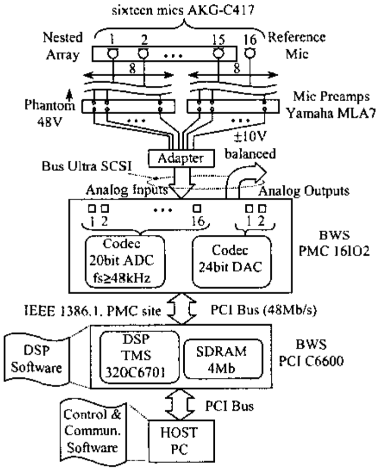

Figure 3: Real-time implementation of the ANS Processor

The real-time array processor uses 16 omnidirectional prepolarized phantom powered AKG-C417 microphones (15 for the array and 1 as clean reference) connected to two 8 channel Yamaha MLA7 preamplifiers, whose outputs act as inputs to a 16 input audio channel Blue Wave Systems (BWS) PMC16IO2 board. Its two 24 bit audio outputs are used as output of the array processor. This PMC board is mounted over the DSP board, a BWS PCI C6600 that includes 4MB of SDRAM plus the Texas

Instruments TMS320C6701 floating point DSP, allowing speeds close to 1 GFLOP. The code of the processor has been developed in ANSI C with Code Composer plus BWS drivers. Lots of specific DSP mathematical and signal processing libraries were also necessary for performance optimization, allowing actually to perform more than tree thousand 512-point-FFT per second.

4. OBJECTIVE EVALUATION OF SPEECH ENHANCEMENT USING E-RASTI

When speech perturbation is caused just by noise, objective evaluation can be obtained by means of SNR measurements, with some consideration of noise spectrum audibility using A-Weighting, the Articulation Index (AI) [8], or considering the masking threshold with the Noise to Masking Ratio (NMR) [5]. Although reverberation is easily detectable when is subjectively considered, objective measurements are very complex. The most accepted method to determine the degradation by reverberation in room acoustics is the Speech Transmission Index (STI).

4.1. STI and RASTI as speech quality estimators

Speech Transmission Index (STI) is based on the fact that the spectrum of the speech signal intensity has very low frequency components, called modulation frequencies. These modulation frequencies correspond with the low frequency envelope of the intensity time signal. When the speech signal is disturbed, the modulation amplitude is reduced. The method based on STI calculates the modulation losses considering 7 audio frequency octave bands and 14 modulation frequencies into each audio band that is, 98 modulation losses altogether. In practice, the RApid Speech Transmission Index (RASTI) [9] is used, because it only considers 4 modulation losses at 500Hz octave band and 5 modulation losses at 2kHz octave band. RASTI is calculate by the modulation losses that have suffered a speech-like signal called RASTI-signal, which is composed by two octave bands (500Hz and 2kHz), and whose corresponding intensity envelope has the appropriate speech-like low frequency components.

In this paper an alternative method is proposed to obtain RASTI index, taking into account not the RASTI-signal, but just the speech signal and calculating the modulation losses of the latter when compared with its clean reference. This method has been called Emulated RASTI (E-RASTI).

4.2. E-RASTI evaluation using speech signals

When RASTI method is applied to a conventional speech signal, the problem is that the intensity envelope associated with the signal does not generally have the modulation frequencies that are present in the RASTI-signal. To overcome the trouble the next method called E-RASTI is proposed. Let us consider an utterance of reverberant speech signal, *y(t)*, with enough length.

1.- Speech signal is filtered in two octave bands, centered at 500Hz and 2kHz obtaining $y_5(t)$ and $y_2(t)$.
2.- Speech signal intensities $I_{5,2}(t)$ are calculated by squaring: $I_{5,2}(t)=y_{5,2}^2(t)$.
3.- Intensity signal is Hanning windowed to avoid border effects.
4.- Low frequency intensity spectrum is calculated using the FFT and resulting $I_{5,2}(\omega)$.
5.- Low frequency spectrum is band-pass filtered to obtain nine plus two (for the 0-frequency level) intensity levels.
6.- The intensity values are used to calculate E-RASTI index, according with the general method described in [9].
7.- RASTI index of the noisy-reverberant speech signal is compared with that one associated with the clean speech signal or the processed speech signal, obtaining E-RASTI.

Figure 3 illustrates the method described in 4.2. To obtain the levels at the modulation frequencies it has been necessary a band-pass filtering of the low frequency intensity spectrum.

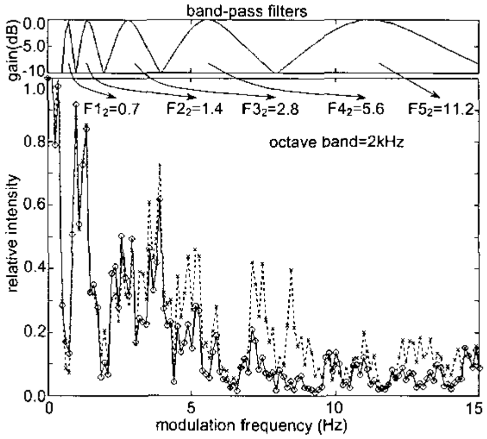

Figure 3: Low frequency intensity spectrum for speech signal. Dashed: processed. Solid: not processed.

5. RESULTS USING THE AUDIBLE NOISE SUPPRESSION (ANS) ARRAY

The performance of the ANS processor has been tested and compared with the MW (Modified Wiener) method obtaining results with two kinds of speech databases called "real database" and "simulated database". "Real database" consists in speech files from the Carnegie Mellon University real multichannel database [2]. It contains simultaneous recordings of a reference signal from a head-mounted close-talk microphone and a 15-microphone array as in figure 1. In this paper the following subcorpora have been used: *arr4A* corresponds to a noisy laboratory with low reverberation and speech source at 1 meter in array axis; *arrC1A*, recorded in a meeting room with high reverberation and low noise (source at 1 meter); *arrC3A*, in the same meeting room but the speech source located at 3 meters in array axis. "Simulated database" consists in speech files with reverberation ($T_{60}\approx 1$s) and random noise artificially added. It has been generated from the close-talk recording of "real database"..

Initially SNR-type objective evaluators have been considered to evaluate the system abilities, that is, the A-Weighting Signal-to-Noise-Ratio (SNR_A), the Articulation Index (AI) and the Noise-to-Masked-Ratio (NMR). In every case the improvement between input and output has been considered. The gain in SNR_A called $GSNR_A$ has been computed as follows:

$$GSNR_A = 10\cdot log\frac{\sum_{k=0}^{k=N-1}\left[\left|Y_{in}(k)-X(k)\right|\cdot A(k)\right]^2}{\sum_{k=0}^{k=N-1}\left[\left|Y_{out}(k)-X(k)\right|\cdot A(k)\right]^2}\ [dB] \quad (7)$$

where k is the frequency index, N the window length, Y_{in} the unprocessed speech spectrum, $Y_{out}(k)$ the processed speech spectrum, $X(k)$ the clean speech and $A(k)$ the A-Weighting filter.

The gain in articulation index (AI) is:

$$GAI = AI_{out} - AI_{in} \qquad (8)$$

using the method described in [8] to obtain AI. The NMR represents an objective evaluator based on the masking threshold and it indicates the noise audibility. The gain in NMR can be calculated by (9),

$$GNMR = 10 \cdot log \frac{\sum_{b=0}^{B-1} \frac{\frac{1}{C_b} \sum_{k=k_{lb}}^{k=k_{hb}} \left| Y_{in}(k) - X(k) \right|^2}{T_b}}{\sum_{b=0}^{B-1} \frac{\frac{1}{C_b} \sum_{k=k_{lb}}^{k=k_{hb}} \left| Y_{out}(k) - X(k) \right|^2}{T_b}} \; [dB] \qquad (9)$$

where b is the index of the critical band, B is the number of critical bands considered, k_{lb} and k_{hb} are respectively the lower and upper frequency indexes associated with critical band b and C_b is the number of bins from critical band with index b.

In the table 1, it has been only considered frames with speech activity, averaging 15 speech utterances for "real database" and 100 utterances for "simulated database".

subcorpus	GSNRA(dB)		GAI		GNMR(dB)	
	ANS-meth.	MW-meth.	ANS-meth.	MW-meth.	ANS-meth.	MW-meth.
arr4A	2.0	0.9	0.07	0.020	6.7	3.1
arrC1A	1.0	0.4	0.03	-0.002	2.1	1.1
arrC3A	0.6	0.5	0.01	-0.002	1.8	1.3

(a)

Input SNR (dB)	GSNRA(dB)		GAI		GNMR(dB)	
	ANS-meth.	MW-meth.	ANS-meth.	MW-meth.	ANS-meth.	MW-meth.
0	9.6	6.0	0.20	0.09	18.3	8.1
10	6.5	5.4	0.19	0.11	12.6	7.6

(b)

Table 1: Results with objective SNR-type evaluators
(a) "Real database" (b) "Simulated database"

The above results show better performance for the ANS processor when compared with the MW processor in all tested conditions with all enhancement estimators (GSNRA, GAI and GNMR), both with real (1.a) and simulated data (1.b). In table 2, E-RASTI indexes are shown obtained with different processing configurations. In order to have better comparison with other systems our All Pass-Minimum Phase (AP-MP) array [2], which obtains excellent dereverberation but limited abilities with non-coherent noise, has also been included in this comparison. In this table the input is the processed (or, in the last column, original) speech and the output is always a noisy or reverberant signal. Consequently the lower the obtained E-RASTI the better the difference between input and output and so the better will be the processor performance.

subcorpus	in: AP-MP method out: noisy signal	in: ANS method out: noisy signal	in: MW method out: noisy signal	in: original signal out: noisy signal
arr4A	0.83	0.92	0.94	0.83
arrC1A	0.78	0.91	0.91	0.83
arrC3A	0.79	0.87	0.85	0.78

(a)

Input SNR (dB)	in: AP-MP method out: noisy signal	in: ANS method out: noisy signal	in: MW method out: noisy signal	in: original signal out: noisy signal
0	-	0.46	0.56	0.51
10	-	0.73	0.75	0.73
> 25	0.77	-	-	0.80

(b)

Table 2: E-RASTI values between input and output
(a) "Real database" (b) "Simulated database"

In the table (2-a), the proposed ANS array performs better than MW, but also AP-MP performs better than ANS. However, AP-MP performs properly just with SNR greater than 15 dB, which is shown in the table (2-b), so the best global performance in both reverberant (*arrC1A, arrC3A*) and noisy (*arr4A, SNR=10dB, SNR=0dB*) environments is obtained by the proposed ANS array processor.

Additionally, in the table (2-b), as SNR decreases E-RASTI also decreases, so a good performance of the array processors and the E-RASTI estimator is shown.

6. CONCLUSIONS

In this paper a real-time microphone array, based on audible noise suppression, has been implemented, tested and compared with the modified Wiener filtering method in the same conditions. The proposed ANS system obtains better objective but especially subjective results with different levels of ambient noise and different reverberation amount, especially in low SNR conditions. Additionally, the system has been implemented with a TMS320C6701-based system, running all the described processes in real time with about 20 to 30% of available time-computing resources.

Also, E-RASTI intelligibility index has been proposed and applied to noisy and reverberant signals. Results have shown its efficiency in reverberation detection and therefore it can be used as objective estimator of the reverberation level present in reverberant speech.

Future work considers the inclusion of dereverberation abilities of the All Pass-Minimum Phase system [2][3] into the Audible Noise Suppression array processor, which has already being performed in high level simulations but remains to be objectively evaluated and encapsulated into our DSP-based system.

7. REFERENCES

[1] Boll, S.F., "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on Speech and Audio Processing,* vol. ASSP-27, pp.113-120, 1979.

[2] Sánchez-Bote, J.L., González-Rodriguez, J., Ortega-García, J., "A new Approach to dereverberation and noise reduction with microphone arrays", *Proc. EUSIPCO,* pp.183-6, 2000.

[3] *González-Rodriguez J., Sánchez-Bote J.L. and Ortega-Garcia, J.,* "Speech dereverberation and noise reduction with a combined microphone array approach", *Proc. ICASSP,* pp.1037-40, 2000.

[4] Akbari, A., Le-Bouquin, R., Faucon, G., "Optimizing speech enhancement by exploiting masking properties of the human ear", *Proc. ICASSP,* pp.800-3, 1995.

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

[6] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. on Speech and Audio Proc.,* vol.7, no.2, pp.126-37, 1999.

[7] Johnston, J.D., "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on Selected Areas in Comm.* vol.6, no.2, pp.314-23, 1988.

[8] Kryter, K., " Methods for the calculation and use of the articulation index", *J. Acous. Soc. Am.* vol.34, p.1689, 1962.

[9] Steeneken, H.J.M. and Houtgas, T., "Rasti: a tool for evaluating auditoria", *Brüel & Kjaer Tech. Rev.,* no.3, 1985