# A New Auditory Based Microphone Array and Objective Evaluation Using E-RASTI

*J.L. Sánchez-Bote, J. González-Rodríguez, and D. Simón-Zorita*

Speech and Signal Processing Group (ATVS)
Departamento de Ingeniería Audiovisual y Comunicaciones (DIAC)
E.U.I.T. Telecomunicación - Universidad Politécnica de Madrid
Ctra. Valencia, km. 7 - Campus Sur, 28031 Madrid, Spain
jbote@diac.upm.es - http://www.atvs.diac.upm.es

## Abstract

Two are the goals of the work presented in this paper. The first one is the implementation of a new method of speech enhancement using microphone arrays. This method gets noise reduction of speech signal using the masking properties of the human auditory system. The second goal of the paper is to use RASTI index (RApid Speech Transmission Index) for objective evaluation of speech signal quality through E-RASTI evaluation. What is new is that E-RASTI is applied to speech signals and no to RASTI-like signals. The E-RASTI index is specially suited to test reverberant speech and has been used here to evaluate the reverberation reduction produced by a microphone array based on all-pass and minimum-phase decomposition or multichannel liftering. Noise reduction evaluation has been performed with the E-RASTI index and also with more traditional methods, based on Signal to Noise Ratios (SNR). Results have demonstrated the good performance of the noise suppressor and the E-RASTI objective quality evaluator.

## 1. Introduction

The degradation sources that can be found in a recorded speech signal are noise and reverberation. The different nature of these two perturbations makes necessary to tackle the problem with different schemes. The most recent methods to enhance speech contaminated by noise are based on the assumption that signal and noise are fully uncorrelated. Noise components can be eliminated by filtering the short-time spectral amplitude of the degraded signal. Two ways to do this have been proposed, by means of spectral subtraction [1], or Wiener filtering [2]-[3]. However later works in single channel speech enhancement, have tested the good performance of noise filtering by using the masking properties of the human auditory system [4]-[6]. This technique is known as Audible Noise Suppression (ANS) and is based on processing only those noise amounts that are over the subjective audible threshold, which is evaluated in each critical band. The ANS method improves subjective perception of the residual noise resulting from that processing, even though objective considerations with SNR measurements show no improvements. The contribution of this work is to use a multichannel system, which can make better estimations for the masking thresholds of the noise-free speech signal. Additionally, this paper introduces a new method called E-RASTI and 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 and has been specially applied to test the dereverberation skills of a multichannel liftering based processor [2]-[3].

## 2. Multichannel speech enhancement using the masking properties of the human auditory system

We are dealing with the problem of improving speech signal contaminated by ambience noise under conditions of very low SNR. All systems that have been used recently are based on some form of noisy speech filtering, in accordance with the estimated SNR. If noise is very high it is necessary to highly reduce the speech levels at particular frequency components, producing and unacceptable loss of subjective quality. Many solutions have been proposed to make more comfortable the residual noise inherent to this processing, however it seems that further improvements are needed. In this paper we have used the masking thresholds [7] of the human auditory system to control the amount of noise reduction that is necessary in each time frame. The philosophy of the method is that noise only should be reduced until the called masking threshold of the speech signal which is related with the level of the noise-free speech in each critical band. If the noise level underpasses the masking threshold it may be considered as subjectively not audible. The masking threshold obtaining process is described in [7]. When the masking threshold is established, and this is not a trivial matter, it is necessary to filter the noisy speech as follows. When the masking threshold is low, high noise suppression is needed, and vice versa. Many suppression functions using the masking threshold may be applied. In this work we have chosen a nonlinear function as described in [5].

### 2.1. Estimation of noise-free speech signal with microphone arrays

The masking thresholds must be calculated from clean speech. In a real system the noise-free speech is unavailable and it must be estimated. In this paper we have used Wiener filtering with coherence modification [2]. This method called Modified Wiener method (MW) has been implemented in a microphone array picking up system. The main advantage of this configuration is that by using the multichannel information both coherent and non-coherent noise can be detected.

The gain of this modified Wiener filter is:

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

with

$$C(\omega)=\frac{G_{x0}(\omega)}{\sqrt{G_{x\,x}(\omega)G_{00}(\omega)}} \quad (2)$$

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 last one just considering temporal frames with no speech activity, $G_{xx}(\omega)$ is the estimation of noisy speech autospectrum obtained by beamforming all channels in three frequency subbands, and $\alpha$ 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 central channel of the array or reference channel and the autospectrum of the reference channel.

We use expression (1) to obtain a noise-free speech estimator as is established in the next equation:

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

where $\hat{S}_{MW}(\omega)$ is the estimation of the clean speech spectrum from Wiener filtering and $Y(\omega)$ is the noisy speech spectrum. $\hat{S}_{MW}(\omega)$ 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 $\hat{S}_{ANS}(\omega)$ can be obtained with:

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

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

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

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

The masking threshold is represented by $T(\omega)$ and must be obtained from clean speech [estimated in $\hat{S}_{W}(\omega)$] using the method described in [7]. The parameter $T(\omega)$ must modify $a(\omega)$ as follows: when the threshold were 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 last:

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

with $N(\omega)$ as the noise autospectrum which can be estimated in the no 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. Objective evaluation of speech enhancement using E-RASTI

The great difficulties inherent to objective evaluation of speech quality are well known. When the speech perturbation is caused by noise the objective evaluation may be easier. For example by means of SNR measurements with some consideration of noise spectrum audibility, using A-Weighting or Articulation Index, AI [8], or even considering the masking threshold with the Noise to Masking Ratio NMR [5]. But usually the biggest problem appears when testing the reverberation reduction. Although reverberation is easily detectable when is subjectively considered, to make objective measurements about this matter is usually very problematic. Nowadays, the more accepted method to determine the degradation by reverberation is the Speech Transmission Index (STI), developed in its origin to evaluate intelligibility.

### 3.1. STI and RASTI as speech quality evaluators

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 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 more much used than the former, because it only considers 4 modulation losses at 500Hz octave band and 5 modulation losses at 2kHz octave band, in total 9 modulation losses. The practical method to implement RASTI method consists in analyzing the modulation losses that has suffered a speech-like signal when is perturbed by reverberation or noise. This signal is called RASTI-signal, and is composed by two octave bands (500Hz and 2kHz), whose corresponding intensity envelope has the appropriate speech-like low frequency components.

In last years, it has been shown that the method based on RASTI-signal is very suitable to test intelligibility losses, and so, we consider that it can be used also to evaluate quality losses of speech signal, specially when this one is distorted with reverberation.

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 last when is compared with original speech signal. This method has been called Emulated RASTI (E-RASTI).

### 3.2. 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.

1.- Let us consider an utterance of reverberant speech signal, $y(t)$, with enough time length. That is, it contains complete sentences, in such a manner that the low frequency envelope can be appreciated. If necessary, some kind of fitting must be done. For example, clearing excessive signal blanks at the

beginning or the end of the utterance that can modify the natural speech modulation. If needed (for example because the speech signal is too short), a time lengthening must be done, normally by repeating *y(t)* to obtain an about 8 second length frame.

2.- The speech signal is filtered in two octave bands, centered at 500Hz and 2kHz obtaining $y_5(t)$ and $y_2(t)$.

3.- Next, the speech signal intensities $I_{5,2}(t)$ are calculated by squaring: $I_{5,2}(t)=y_{5,2}^2(t)$.

4.- Then the intensity signal is windowed to avoid border effects if the frame is suddenly ended. The hanning window fits well here.

5.- The low frequency intensity spectrum is calculated using the FFT and resulting $I_{5,2}(\omega)$.

6.- The low frequency spectrum is band-pass filtered to obtain nine plus two (for the 0-frequency level) intensity levels, according to the frequencies of table 1.

7.- The intensity values are used to calculate E-RASTI index, according with the general method described in [9].

8.- The RASTI of the noisy-reverberant speech signal is compared with that one associated with the clean speech signal or the processed speech signal and obtaining E-RASTI.

| octave band=500Hz | | | | | |
|---|---|---|---|---|---|
| Modulation freq. (Hz) | $F0_5$=0 | $F1_5$=1 | $F2_5$=2 | $F3_5$=4 | $F4_5$=8 |
| **octave band=2kHz** | | | | | |
| Modulation freq. (Hz) | $F0_2$=0 | $F1_2$=0.7 | $F2_2$=1.4 | $F3_2$=2.8 | $F4_2$=5.6 | $F5_2$=11.2 |

*Table 1:* RASTI *m*odulation frequencies

## 4. System description and speech databases

The microphone array arrangement used to achieve speech enhancement is shown in figure 1.

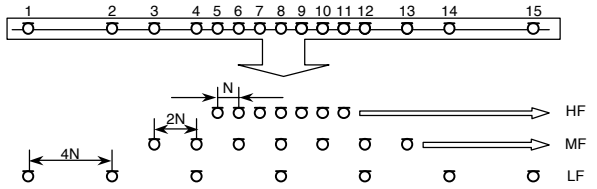


*Figure 1:* Nested array of fifteen microphones

Speech signal spectrum is split into three frequency subbands, each one from a microphone group, as shown in figure 1. In figure 2 the processor scheme is presented. As can be seen, the multichannel signal may be processed in two ways, by ANS or MW methods.

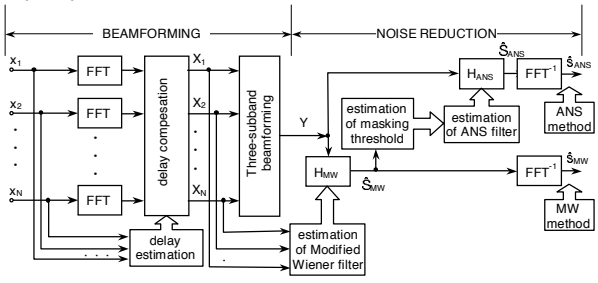


*Figure 2:* ANS method processor

To process dereverberation, we have used the system described in [2] that uses multichannel liftering based on All-Pass and Minimum-Phase decomposition (AP-MP method). Although in the final system both ANS and AP-MP methods are intended to be implemented together, in this paper these two methods are tested separately.

Results have been obtained 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 has simultaneous recordings of a reference signal from a head-mounted close-talk microphone and a 15-microphone array with the same configuration of figure 1. The database contains 10 subcorpora, taken in different rooms and ambient conditions. 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$1s) and random noise artificially added. It is based on the same array configuration and has been implemented using as reference the close-talk recording of "real database".

## 5. Results using ANS method

The performance of the ANS processor has been tested and compared with the MW method, using both "real" and "simulated" databases. In a first stage SNR-type objective evaluators have been considered to evaluate the system skills, 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 is:

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

using the method described in [8] to obtain AI. The NMR represents an objective evaluator based on the masking threshold and it indicates if the noise is audible or not. The gain in NMR can be calculated by (9),

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

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 $b$.

All results have been obtained considering only he frames with speech activity, and are shown in table 2.

Figure 3 illustrates the method described in 3.2 to obtain

E-RASTI from speech signal. As explained before, to obtain the levels at the modulation frequencies it has been necessary a band-pass filtering of low frequency intensity spectrum.

| 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 2:* Results with objective SNR-type evaluators. (a) "Real database". (b) "Simulated database".

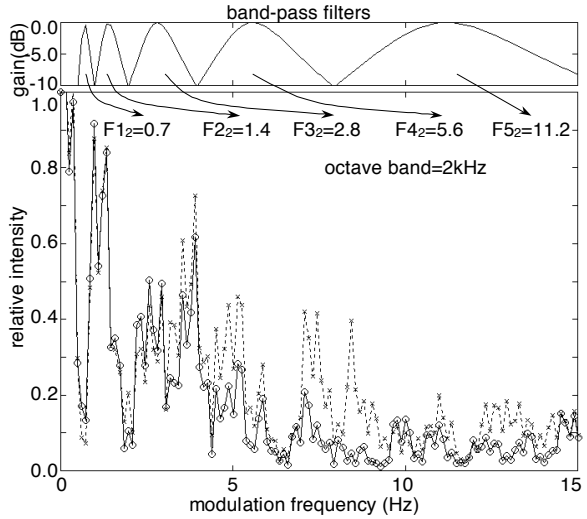


*Figure 3:* Low frequency intensity spectrum for speech signal. Dashed: processed. Solid: not processed.

| 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 |
| **∞** | 0.77 | - | - | 0.80 |

**(b)**

*Table 3:* E-RASTI values between input and output. (a) "Real database". (b) "Simulated database".

Table 3 contains E-RASTI indexes obtained with different processing configurations. In this table the input is the processed (or 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. For example, in table (3-b) choosing the same column, when SNR is getting lower E-RASTI decreases and so the noise at the input will be lower. The efficiency of the AP-MP method in reverberation reduction can be seen in tests with low noise speech. For example, looking at the subcorpora *arrC1A* and *arrC3A,* the E-RASTI index for AP-MP method is lower than that corresponding to ANS and MW methods. This is because in low noise situations, the noise suppression systems have bad performance.

All shown results have been achieved averaging 15 speech utterances for "real database" and 100 speech utterances for "simulated database".

## 6. Conclusions

In this paper a microphone array, based on audible noise suppression, has been tested and confronted with the Wiener filtering method working in the same conditions. Subjectively talking, the results from the former were better, although the pure SNR improvements of both were similar. The perception of background noise was greater in the second case, specially with low input SNR.

On the other hand, E-RASTI intelligibility index has been proposed and applied to noisy and reverberant signals, processed with the two systems related before and besides with a reverberation suppressor, based on multichannel liftering. Results have shown that E-RASTI can detect great SNR improvements, as is shown in table (3-b) for low SNR, but the great advantage of this method is its efficiency in reverberation detection and therefore it can be used for objective evaluation of reverberant speech. Results have agreed the use of E-RASTI index for dereverberation evaluation. However we have found some problems with some speech utterances with low modulation in its low frequency intensity spectrum. For near future we propose to work establishing criteria to make speech signal suitable for E-RASTI measurements in any condition.

## 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-Rodríguez, J., Ortega-García, J., "A new Approach to dereverberation and noise reduction with microphone arrays", *Proc. EUSIPCO*, pp.183-186, 2000.

[3] González-Rodríguez J., Sánchez-Bote J.L. and Ortega-García, 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 Transactions on Speech and Audio Processing*, 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. Acoustical Soc. Am.* vol.34, p.1689, 1962.

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