SOURCE SEPARATION FOR WFS ACOUSTIC OPENING APPLICATIONS

J. A. Beracoechea, S. Torres-Guijarro, L. Ortiz-Berenguer, F. J. Casajús

Departamento de Señales, Sistemas y Radiocomunicaciones
Universidad Politécnica de Madrid
berako@gaps.ssr.upm.es

ABSTRACT

This paper proposes a new scheme to reduce coding bit rate in array based multichannel audio applications like the acoustic opening, which can be used for modern teleconference systems. The combination of beamforming techniques for source separation and wave field synthesis allows a significant coding bit rate reduction. To evaluate the quality of this new scheme, both objective and subjective tests have been carried out. The objective measurement system is based on the Perceptual Audio Quality Measure of the binaural signal that the listener would perceive in a real environment.

1. INTRODUCTION

Over the last years there has been a significant development of multichannel audio. These technologies are evolving towards systems capable of recreating true 3D audio fields in a listening area as wide as possible. This evolution implies rising the number of loudspeakers used for sound reproduction. Today it is possible to find commercial products that use 5.1 channels but new systems will use many more channels to increase the perceived spatial sensations. If a high number of channels are involved in an audio system, this means that a large amount of audio material needs to be transmitted or recorded. This justifies recent attention on multichannel audio coding systems, that try to reduce the overall bit rate without penalizing quality.

One of the most promising multichannel audio systems is Wave Field Synthesis (WFS). Wave Field Synthesis technique reproduces an acoustic field inside a volume from the signals recorded or computed on a given surface. It is based on the Huygens principle. According to this principle, the propagation of a wave through a medium can be qualitatively described by adding the contributions of all secondary sources positioned along a wave front [1]. This means that if we know the wave field on the boundary surface S of a closed, source-free volume V it is possible to know the sound pressure in any point within that volume. From a practical point of view, this means that if we cover a plane with an array of omni directional loudspeakers, being driven with signals corresponding to the normal velocity distribution in that plane generated by virtual source, a spatially correct wave field of a point positioned behind the array is synthesized.

The application of microphone and loudspeaker array systems to enhance perceived sensations is under study. Using a high number of microphones (more than 20) in a linear array makes possible to sample the entire acoustic field. This field can be recreated in another location by means of Wavefield synthesis [2] using a loudspeaker array.

We will focus on a audio communication system known as acoustic opening [3]. Arrays of transducers are used to produce the illusion that there is a mechanical opening between two remote rooms. A simplified system using an array of microphones and array of loudspeakers can be seen in Figure 1 and Figure 2. The acoustic wave field is recorded or sampled using the microphone array, coded, transmitted and reproduced through an array of loudspeakers. This configuration may be used to develop hands-free acoustical human/machine interfaces (teleconference systems). As can be seen in [3], if the number of transducers is sufficiently high the illusion would be perfect. Unfortunately, working with so many channels means that there is an enormous amount of information to deal with.

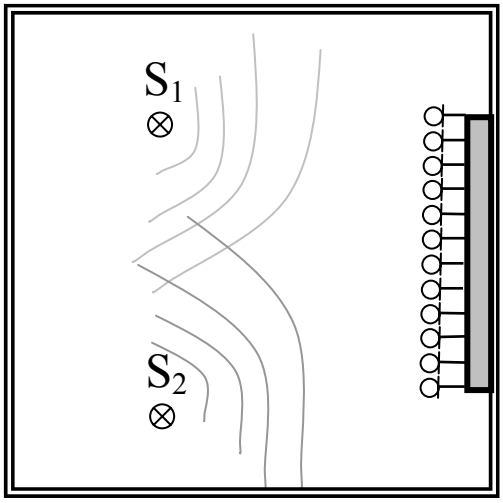

Figure 1: *Dry sources sampled by an array of microphones*

2. CODING APPROACHES

The two most used multi-channel codification systems are Dolby AC-3 and MPEG Advanced Audio Coding (AAC). AC-3 is the audio standard chosen for high-resolution television (HDTV), and it is able to compress 5.1 audio signals using 384 kbits/s. AAC is at the moment the most powerful multi-channel codification system within the family of MPEG coders. It is able to compress 5.1 audio signal using 320 kbits/sec without apparent loss of quality. Both schemes use perceptual models to hide coding distortions. Although they are very powerful systems that support a high

number of channels, they are optimized to encode 5.1 recordings. Thus, the multichannel strategies employed (Mid/Sum Coding and Intensity Coding) try to exploit the correlation between symmetric channel pairs (e.g. L-R and Ls-Rs), but are unable to eliminate the existing correlation among the rest of the channels.

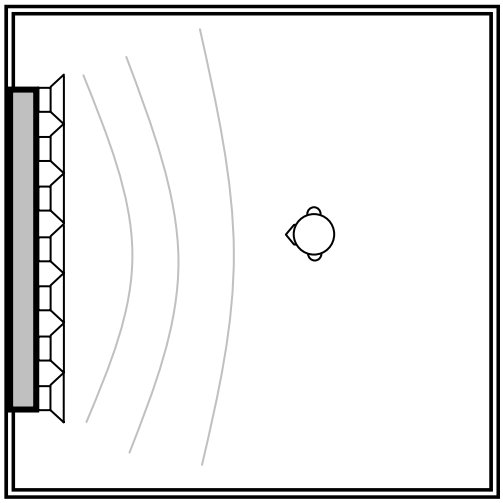

Figure 2: *The loudspeaker array re-builds the acoustic field*

One of the possibilities to decrease this bit rate employs the Karhunen-Loeve Transform (KLT) [4] to improve the overall behaviour of the system. In this approach, decorrelation of the whole multichannel signal is achieved by means of the KLT. Once decorrelated, the audio channels are independently processed by a bank of perceptual codecs as we can see in Figure 3. Codification bit rate is distributed among them depending on the energy of each decorrelated channel. The exact distribution is adjusted to obtain the best final quality for a given total rate. As we can see in [5] this new approach allows a 20%-50% bit rate reduction depending on the nature of the multichannel signal.

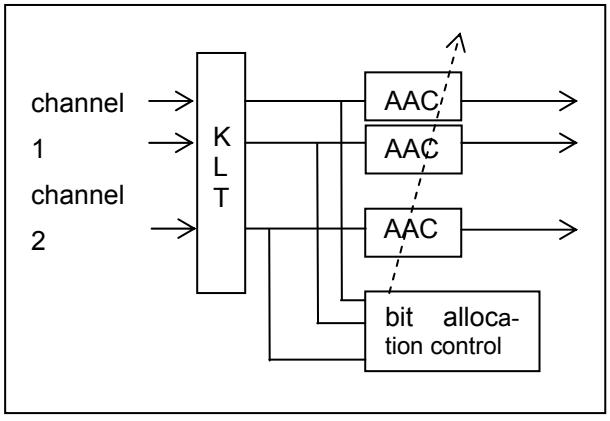

Figure 3: *KLT+AAC multichannel coding*

Despite of this reduction, the number of channels to be transmitted still equals the number of microphones in the array. In order to reduce the number of channels to be transmitted, array processing methods are explored in the following.

3. SOURCE SEPARATION APPROACH

To develop a wave field synthesis approach we have two possibilities: in the first one, all microphone signals are transmitted to the same number of loudspeakers at the receiving end. This system receives the name of 'hard-wired wave field transmission system'. In this approach the compression gain comes from exploiting the correlation between channels (as seen in previous Sections). However there is another possibility where the signals that feed the secondary sources (the loudspeakers at the receiving room) are extrapolated of a enough dense set of measured impulse responses. This new approach has a tremendous impact from a codification point of view. Now, it is possible to send only the dry sources and the impulse responses of the room and recreate the wave field at reception. This leads us to the problem of obtaining the dry sources given that we only know the signals that the microphone array captured. Basically, this is a source separation problem. In Figure 4 we can see the full scheme where a WFS system synthesizes the wave field produced by primary sources in the simulated room.

From a mathematical point of view, the problem to solve can be resumed in expressions (1), (2) and (3). There are P statistically independent wideband sound sources ($S_1 ... S_P$) in a M-microphone room ($P<M$). Each microphone signal is produced as a sum of convolutions between sources and H_{ij} , which represents a matrix of z-transfer functions between P sources and M microphones. This transfer function set contains information about the room impulse response and the microphone response. The number of sources (M) is always lower than the number of microphones (P). We have:

$$\begin{vmatrix} X_1(z) \\ X_2(z) \\ \vdots \\ X_M(z) \end{vmatrix} = \begin{vmatrix} H_{11}(z) & \dots & H_{1P}(z) \\ H_{21}(z) & \dots & H_{2P}(z) \\ \vdots & \vdots & \vdots \\ H_{M1}(z) & \dots & H_{MP}(z) \end{vmatrix} \begin{vmatrix} S_1(z) \\ S_2(z) \\ \vdots \\ S_P(z) \end{vmatrix} \tag{1}$$

$$\mathbf{X} = \mathbf{HS} \tag{2}$$

We make the assumption that source signals S are statistically independent processes, (which is a sufficient condition for source separation) so the minimum generating signals $\mathbf{\Gamma}$ will be the same as the number of sources P. We need $\mathbf{\Gamma}$ to be as similar as possible to $\mathbf{S}$ (original dry signals). Ideally $\mathbf{J}$ would be the pseudoinverse of $\mathbf{H}$, however we may not know the exact parameterization of $\mathbf{H}$. In the real world spatial separation of sources from an output of a sensor array is achieved using beamforming techniques.
Thus, we let

$$\mathbf{\Gamma} = \mathbf{JHS} \tag{3}$$

4. BEAMFORMING: GENERALIZED SIDELOBE CANCELLER

For acoustic openings, microphone arrays together with robust adaptive beamforming techniques allow the extraction of desired signals from many kind of interferers (background noise, reverberation, or competing talkers). One of the most used beamforming algorithms is the Generalized Sidelobe Canceller [6] that can obtain a high interference reduction performance with a small number of microphones arranged in a small space.

One of the biggest concerns in using GSC is that we need to know the direction of arrival (DOA) of the primary source. That means that we need to know quite accurately the position of the speakers in the room. This can be achieved using DOA-determination algorithms, like the MUSIC algorithm [7]. The MUSIC algorithm was developed by Schmidt to determine direction-of arrival angles for multiple sources and although it offers good results if the primary sources are narrow band signals, with broadband signals (like voice) the results are not so good [8]. The resolution of this problem is beyond the objectives of this article, for our work we suppose the DOA is known.

We can see the general layout of the GSC in Figure 5. First of all, the microphone signals are time delayed steered ($\tau_1,\ldots,\tau_M$) to produce signals which ideally have the desired signal in phase with each other. If we add all these signals ($d(n)$) we have a classical delay and sum beamformer. Usage of a simple delay-and-sum beam former leads to target signal cancellation at high frequencies so we need to complicate the system with an adaptive algorithm to improve the overall performance. A delayed version of $d(n)$: $d'(n)$ (to keep causality) is used as reference for the adaptive sidelobe canceling path. [8]. Depending on how precise is the DOA information this reference would be good enough. The delayed signals (before adding) are then sent to the blocking matrix. The purpose of the blocking matrix is to block out the desired signal from the lower part of the GSC. The idea is to adaptively cancel out noise and interference sources, therefore we only want noise to go into the adaptive filters FIR1… FIR(M–1). At the present moment we have used a very simple blocking matrix, which means that the outputs of the matrix are the difference between successive signal samples:

$$B = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{4}$$

If the input signals of the adaptive filters (FIR1- FIR(M-1)) contain only interferences the multiple input canceller rejects the interferences and extracts the target signal (e(n)). If the target signal leaks through the blocking matrix the adaptive algorithm (we use nLMS for simplicity) cancels both the interference and the desired signal (the original dry source we want to recover). This leakage can be caused by two different causes. First of all, bad DOA tracking; second, a highly reverberant room. If we are working with a highly reverberant room (high T60) some of the reflections of the interference signals may leak into the main lobe. In this case even with highly complex blocking matrixes [9], [10] it may be very difficult to obtain target-free signals from the output of the BM. As we know the exact DOA, the current BM configuration is enough to extract a signal which is quite similar to the original dry source.

5. EXPERIMENTS AND CONCLUSIONS

To obtain the microphone signals we are employing the impulse-response recordings of the varecoic chamber in Bell Labs [11], corresponding to different audio source locations in the chamber and a 22-microphone linear array. This arrangement is perfect for studying the "virtual acoustic opening". We have also considered free field propagation conditions (no chamber) to compare the effects of reverberation. Two different speakers (male and female), which act as primary sources, are placed on both sides of the room. These two signals are convolved with the impulse response of the chamber corresponding to those concrete locations to obtain the 22 microphone signals. In the case of employing free field conditions we only delay and scale properly the signals.

We have considered two different scenarios. In the first one we suppose that we are using a hard-wired WFS system. That means that we 'send' the full 22 channels and the sound field is reconstructed using directly the transmitted signals In the second scenario we apply the beamforming algorithm to recover both original dry signals. In this occasion we suppose that we are only 'sending' these two signals. At reception we rebuild the acoustic field using WFS techniques. The results are based in the comparison of these two scenarios.

For this comparison we have used objective and subjective tests. One of the tools to be used is based on the ITU standard Perceptual Audio Quality: PEAQ [12] which is widely accepted for codec comparison. We have specifically used an implementation if the basic version of the recommendation ITU-R BS. 1387-1 [13]. This module measures the perceptual difference between the original and processed signal by means of the so called Objective Difference Grade (ODG). The output of the module is a figure between 0 and –4 where 0 means "no perceptible degradation" and –4 means "very annoying degradation"

PEAQ was developed for evaluating the quality of mono of stereo signals, not for multichannel audio. The proposed solution to this problem is to synthesise the binaural signal which is computed with the following guidelines.

- In the first scenario, we obtain the loudspeaker driving signals supposing that we are using a hard –wired WFS systems as if we had sent the 22 channels
- In the second one we obtain the pseudo-dry sources using the beamforming algorithm and reconstruct the loudspeaker driving signals using WFS.
- Loudspeakers are considered ideal; in a real non simulated environment. Equalization should be implemented
- Free field propagation from the loudspeakers to the listener is assumed
- To obtain a true binaural signal, the effect of external ear, head, shoulders, etc. is taken into account by using HRIRs (Head Related Impulse Responses). The signal coming from each loudspeaker is filtered with the HRIR that corresponds to that direction of arrival. The particular impulse response set is the one measured with KEMAR (Knowles Electronic Mannequin for Acoustic Research [14]) with diffuse equalization. The listener is positioned in the middle of the reception room. With a true mechanical opening, he would hear the male voice coming from the right and the female voice coming from the left.

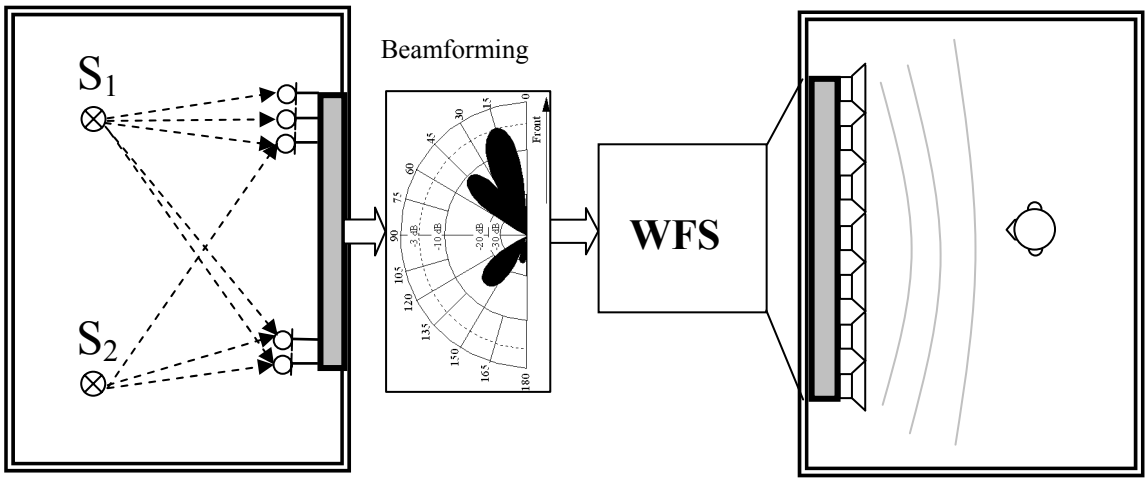

Figure 4: *Acoustical opening coding system.*

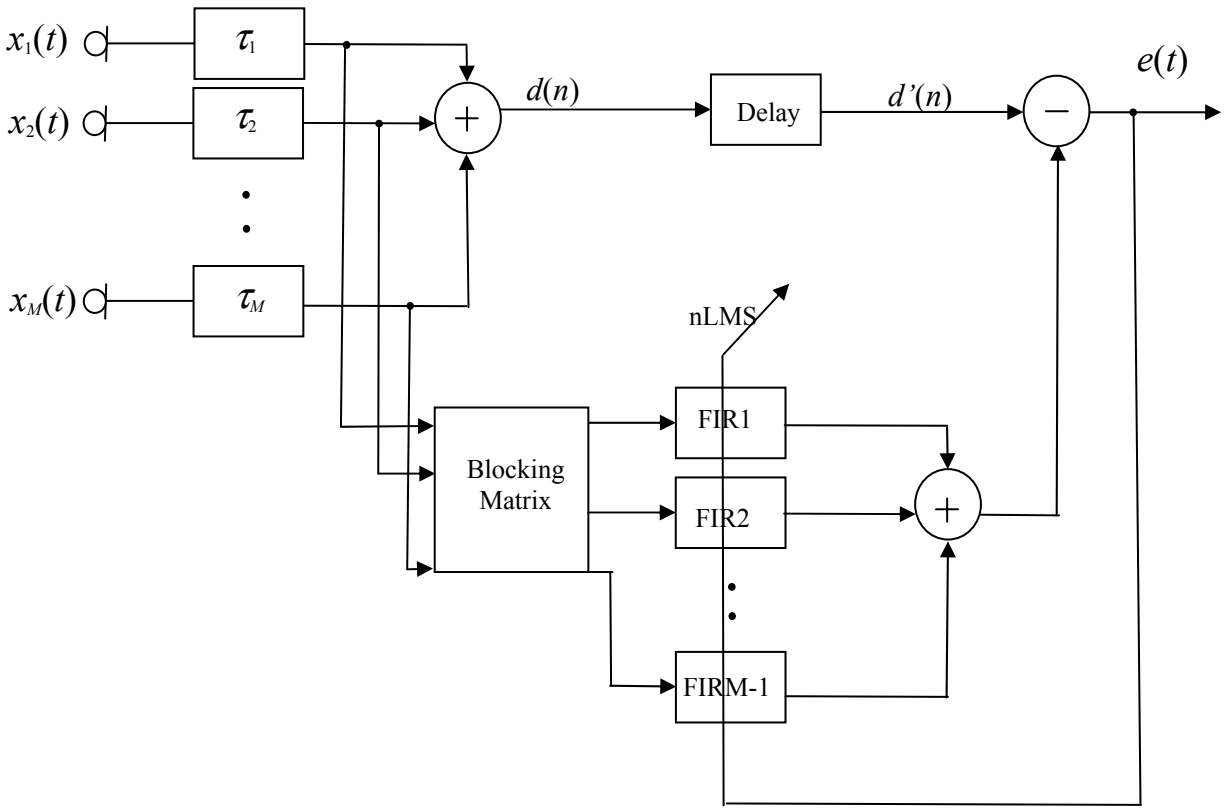

Figure 5: *Generalized sidelobe canceller.*

If we use free field conditions the results are quite promising. The source separation is nearly perfect and the ODG value obtained is -0.7 which means that the distortions are nearly inaudible. This has a great impact from a coding point of view. In the first scenario we 'sent' 22 channels while in the second we only 'sent' two. The bit rate reduction is huge.

Problems arise when using the impulse responses of the varecoic chamber. Due to reverberation, source separation is not so good and you can hear an attenuated version of the female speaker signal in the separated male speaker signal and vice-versa. Also, the performance of PEAQ algorithm is not fully reliable in these conditions. However informal subjective tests with 6 listeners have showed that there is not a big difference between both scenarios in terms of quality at reception (after WFS and binauralization).The intelligibility is even better in the second scenario due to the reduction of the reverberation effect by the adaptive algorithm The spatial sensations are also preserved (we still hear the male voice coming from the right and the female voice coming from the left) which is an important feature. We have noticed that the adaptive algorithm embedded in the sidelobe canceller performs much better when the interferer is white noise instead of a second speaker. The silences between words cause the nLMS algorithm to diverge. In the future it may be necessary to implement some kind of vocal activity detector to stop the nLMS adaptation algorithm in the silences. If we consider noise as the interferer, the source separation becomes a noise cancellation problem. For this case, we can see the behaviour of the GSC in Figure 6.

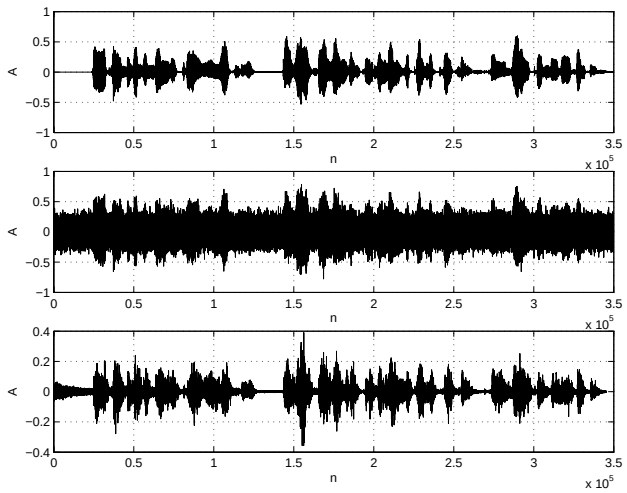

Figure 6: *Noise canceller behavior.*

On the upper part of the picture we can see the original dry signal (male speech). In the middle, the signal recorded in the central microphone of the array (signal + noise). In the lower part we can see the GSC output (pseudo-dry signal). As we can see the noise reduction is quite effective. Taking a closer look at the first samples, you can notice a progressive noise reduction due to the nLMS convergence time. Playing with the adaptation step makes possible to decrease this convergence time. However, in this case, the final SNR would be higher.

The results presented in this paper are still preliminary but we think that are quite promising. Using beamforming together with Wavefield Synthesis to recreate a mechanical acoustic opening may provide us with a very useful tool to drastically reduce the number of channels to transmit (and consequently the bit rate). There still are problems to solve, like the DOA estimation, the effect of high reverberant rooms and the development of better quality measures, but it seems that the path is correct and that future teleconference systems may benefit from this approach.

6. REFERENCES

[1] Marinus M. Boone, "Acoustic rendering with wave-field synthesis," *ACM Siggraph and Eurographics Campfire*, May 2001.

[2] D. de Vries, Marinus M-Boone, "Wave field synthesis and analysis using array technology," *Proc. 1999 IEEE Workshop of Application of Signal Processing to Audio and Acoustics,* New York, Oct. 1999.

[3] Aki Härmä, "Coding Principles for Virtual Acoustic Openings," *AES 22nd Int. Conf.*, pp. 159-165.

[4] S. Torres, J. A. Beracoechea, F. J. Casajús, L. Ortiz, "Multichannel audio decorrelation for coding," *Proc. of Int. Digital Audio Effects Conf. (DAFX-03),* London, U.K., Sept. 2003.

[5] S. Torres, J. A Beracoechea, F. J. Casajús, I. Perez-Garcia, "Coding strategies and quality measure for multichannel audio," *116th AES Conv.,* Berlin, Germany, May 2004.

[6] L. J. Griffths, C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. AP-30, no. 1, Jan. 1982.

[7] R. O. Schmidt, "Multiple emitter and signal parameter estimation," *Proc. RADC Spectral Estimation Workshop*, pp. 243–258, Oct. 1979.

[8] D. Campbell, *Adaptive Beamforming using a microphone array for hands-free telephony.* Master thesis.

[9] H. Herbordt, W. Kellermann, "Efficient frequency-domain robust generalized sidelobe canceller," *Proc. IEEE Workshop on Multimedia Signal Processing,* Oct. 2001, pp. 377–382.

[10] O. Hoshuyama, A. Sugiyama, A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on Signal Processing,* vol. 47, no. 10, Oct. 1999

[11] Bell Labs varecoic chamber: `http://www.bell-labs.com/org/1133/Research/Acoustics/Varecoic Chamber`

[12] ITU Recommendation BS. 1387 (2002) *Methods for objective measurements of perceived audio quality*

[13] EAQUAL: Implementation of ITU-R recommendation BS. 1387. `http://home.wanadoo.nl/~w.speek/eaqual.htm`

[14] HRTF measurements of KEMAR. `http://sound.media.mit.edu/KEMAR.html`