ACOUSTIC ECHO CANCELLATION FOR SURROUND SOUND USING PERCEPTUALLY MOTIVATED CONVERGENCE ENHANCEMENT

Jürgen Herre[*]

Fraunhofer Institute for Integrated Circuits IIS,
Am Wolfsmantel 33, 91058 Erlangen, Germany
hrr@iis.fraunhofer.de

Herbert Buchner, Walter Kellermann

Multimedia and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{buchner,wk}@LNT.de

ABSTRACT

Acoustic Echo Cancellation (AEC) has become an essential and well-known enabling technology for hands-free communication and human-machine interfaces. AEC for two or more reproduction channels aims at identifying the echo paths between the microphone and each audio reproduction source in order to cancel the associated echo contribution. A number of preprocessing methods have been proposed to decorrelate stereo audio signals in order to enable an unambiguous identification of each echo path and to thus ensure robustness to changing sound source locations. While several of these methods provide enough decorrelation to achieve proper AEC convergence in the stereo case, considerations of subjective sound quality have frequently not been addressed adequately. This paper compares the performance of several methods in terms of both convergence speed and aspects of sound perception, and proposes a novel signal decorrelation approach with attractive properties. The superior performance of the proposed method is demonstrated for 5.1 surround sound reproduction.

Index Terms— Acoustic Echo Cancellation, AEC, Multi-Channel Sound, Surround Sound, Phase Modulation, Perception, Psychoacoustics

1. INTRODUCTION

Acoustic echo cancellation for speech and audio communication where loudspeaker signals feed back into microphones has already been discussed extensively for the single-channel case and for stereo sound reproduction (e.g., [1, 2, 3, 4]). More recently, AEC has been realized for more than two reproduction channels [4, 5] as it is necessary, e.g., for voice-controlled infotainment systems or home theaters using an automatic speech recognizer (ASR). Thereby, multi-channel frequency-domain adaptive filtering (MC-FDAF) algorithms have turned out to be particularly efficient. Figure 1 describes a typical scenario for stereo or multi-channel AEC. From a transmitting room, a sound source (e.g., a speaker) is picked up by P microphones ($P = 2$ for stereo). The microphone signals are transmitted to a receiving room and reproduced via P loudspeakers. At the same time, a microphone in the receiving room picks up speech from a local user. In order to prevent the sound emitted from the loudspeakers coupling into the outgoing microphone signal (which is sent back to the far-end listener or an ASR), AEC attempts to cancel out any contributions of the incoming signals $x_i(k)$ from the outgoing signal by subtracting filtered versions $\hat{y}_i(k)$ of the incoming signals from the outgoing one $y(k)$. This generally requires that cancellation filters (assumed to be length-L FIR filters) are dynamically adjusted by an adaptation algorithm to achieve minimum error signal $e(k)$ and thus optimum cancellation. This is the case when the adaptive cancellation filters

$$\hat{\mathbf{h}}_i(k) = \left[\hat{h}_{i,1}(k), \cdots, \hat{h}_{i,L}(k)\right]^T, \;\; i = 1, 2, \dots, P$$

accurately model the impulse responses $\mathbf{h}_i$ from the emitting speakers to the microphone.

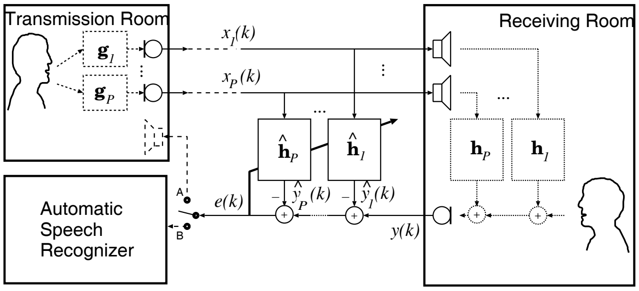

Fig. 1. Scenario for multi-channel AEC.

It has been shown for stereo AEC that a so-called *non-uniqueness problem* exists [6]: If both loudspeaker signals are strongly correlated, then the adaptive filters generally converge to a solution that does not correctly model the transfer functions between the speakers and the microphone, but merely optimizes echo cancellation for the given particular loudspeaker signals. As a consequence, a change in the characteristics of the loudspeaker signals (e.g. due to a change of the geometric position of the sound source in the transmitting room) results in a breakdown of the echo cancellation performance and requires a new adaptation of the cancellation filters. To solve this non-uniqueness problem, various techniques have been proposed to preprocess the signals transmitted from the transmitting room prior to their reproduction in the receiving room in order to *decorrelate* all channels relative to each other and thus avoid this ambiguity. The key requirements for such preprocessing schemes are:

- *Convergence enhancement*: The processing must be able to decorrelate the input signals effectively to ensure rapid and correct AEC filter convergence even for highly correlated / quasi-monophonic audio signals.
- *Subjective sound quality*: Since the preprocessed signals are subsequently reproduced via loudspeakers and listened to by users in the receiving room, the preprocessing must not introduce any objectionable artifacts into the reproduced audio

[*] Work performed at Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Germany.

signals (this may be speech only for hands-free telecommunication applications or any type of audio material including music if used for ASR input enhancement.)

- *Complexity*: Especially for inexpensive consumer equipment, very low computational and memory complexity is desirable.

This paper proposes a novel and flexible approach, based on perceptual considerations which fits these requirements. Moreover, it easily generalizes to the multi-channel case and is demonstrated to be effective in surround sound echo cancellation.

2. KNOWN TWO-CHANNEL PREPROCESSING APPROACHES

A first simple preprocessing method for stereo AEC was proposed by Benesty et al. [7, 8] and achieves signal decorrelation by adding non-linear distortions to the signals. While this approach features extremely low complexity, the introduced distortion products can become quite audible and objectionable, especially for high-quality applications using music signals. Moreover, the generalization of this approach to an arbitrary number of channels is not straightforward.

A second well-known approach consists of adding uncorrelated noise to the signals. In [9], this is achieved by perceptual audio coding / decoding of the signal which introduces uncorrelated quantization distortion that is masked due to the noise shaping according to the coder's psychoacoustic model. A similar effect can be achieved by using a perceptually controlled watermarking scheme, e.g., based on spread spectrum modulation [11]. In both cases the use of an explicit psychoacoustic model plus analysis / synthesis filterbanks is able to prevent audible distortions for arbitrary types of audio signals and may be easily generalized to more than two channels. However, the associated implementation complexity and the introduced delay render this approach unattractive for most applications.

A third approach to AEC preprocessing is to use complementary comb filtering [10] on the two output signals. Unfortunately, this type of processing generally leads to unacceptable degradations of the stereo image perceived by human listeners which make it unsuited for high quality applications.

Still other approaches employ switched / time-varying time-delays [3] or variable all-pass filtering [13] to produce a time-varying phase shift / signal delay between the two channels of a stereo AEC and thus "decorrelate" both signals. Specifically, [3] describes a preprocessing system in which the output signal switches between the original signal and a time-delayed / filtered version of it. As a disadvantage, this switching process may introduce unintended artifacts into the audio signal. [13] describes a system in which an allpass preprocessor is randomly modulating its allpass filter variable. In [14], it was proposed to apply this allpass preprocessor only to the low frequency range up to 1 kHz due to convergence requirements.

3. PROPOSED METHOD

In order to obtain a preprocessing method offering both good decorrelation properties for the enhancement of AEC convergence *and* minimal alteration of the perceived stereo image, the proposed method is based on several considerations. From the previously discussed approaches time-varying modulation of the phase of the audio signal, as proposed in [3, 13], is an effective method which is generally unobtrusive in its perceptual effects on audio signals as compared to other methods while avoiding computationally expensive masking models. Nonetheless, it is difficult to achieve maximum decorrelation while guaranteeing that introducing a time / phase difference between left and right channels does not result in an alteration of the perceived stereo image. Several aspects must be accounted for:

- Interaural phase / time difference is a relevant perceptual parameter for subjective perception of a sound stage [15] and has been used extensively in synthesis of stereo images (e.g. [16]). Consequently, a change in the perceived stereo image can only be avoided if the introduced time / phase difference stays below the threshold of perception, as it applies to audio signals that are reproduced via loudspeakers.
- Optimal AEC convergence enhancement can be achieved if the preprocessing introduces time / phase differences just at the threshold of perception, i.e., applies the full amount of tolerable change.
- As is known from psychoacoustics, the human sensitivity to phase differences is high at low frequencies, and gradually reduces for increasing frequencies, until it fully vanishes for frequencies above ca. 4 kHz.
- Neither a simple time delay modulation nor a low-order time-varying allpass filtering offer the flexibility to tailor the amount of time / phase shifting as a function of frequency, such that the full potential of perceptually tolerable change is exploited.

Hence, in contrast to the earlier phase modulation approaches, the proposed method is designed to allow a perceptually motivated frequency-selective choice of phase modulation parameters (modulation frequency, modulation amplitude, and modulation waveform) by employing analysis / synthesis filterbanks. The input audio signal is decomposed into subband signals by means of an analysis filterbank. Then, the subband phases are modified based on a set of frequency-dependent modulating signals. According to the above considerations, subbands belonging to the low frequency part of an audio signal should be left largely untouched, while subbands corresponding to frequencies above 4 kHz may be modulated heavily. Finally, the modified spectral coefficients are converted back into a time domain representation by a synthesis filterbank. To allow easy access to the signal's phase, a complex-valued filterbank [12] is used, and a phase modification is implemented by a complex multiplication of the subband coefficient with $\mathrm{e}^{\mathrm{j}\varphi(t,s)}$ where $\varphi(t,s)$ denotes the intended time varying phase shift in subband s as discussed below.

3.1. AEC for Stereo Reproduction

Figure 2 shows a simple preprocessor applying the proposed method to stereo signals. The time-varying phase difference between the output signals is produced by a common modulator function $\varphi(t,s)$ which is scaled differently for each subband s, and is applied to both channels in a conjugate complex way, i.e., the phase offset introduced into the left channel has the opposite sign as the phase offset introduced into the right channel signal.

As a consequence of the phase modulation, a frequency modulation is introduced with a frequency shift that is proportional to the temporal derivative of the phase modulation function. Therefore, in order to avoid a perceptible frequency modulation of the output signal, it is preferable to choose a smooth modulating function, such as a sine wave at a relatively low modulation frequency, or a sufficiently smooth random function (similarly to [13]).

As an example for suitable parameters, a CMLT filterbank with a window length of 128 was used in this paper resulting in 64 subbands. For stereo preprocessing, a phase modulation of $\varphi(t,s) = a(s)\sin(2\pi f_{\mathrm{m}} t)$ is applied with a modulation frequency of f_{m} =0.75 Hz. The modulation amplitude $a(s)$ reflects the frequency-dependent perceptual sensitivity to phase modulation in a common acoustic speaker/room/listener setup and has been optimized by a listening

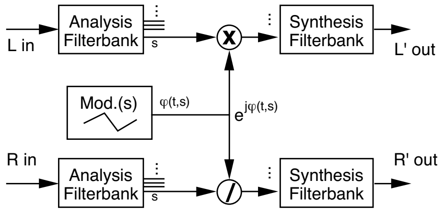

Fig. 2. Preprocessing for a stereo pair of audio channels.

procedure. It scales from 10 degrees at low frequencies to 90 degrees for frequencies at and above 2.5 kHz. Figure 3 depicts the modulation amplitude for the first 12 subbands.

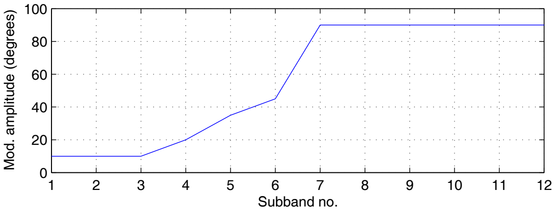

Fig. 3. Phase modulation amplitude as a function of subband.

3.2. AEC for for Multi-Channel / Surround Reproduction

For generalizing the above method from stereo to the multi-channel case, we not only have to generalize the modulation scheme of Sect. 3.1 but also need to consider the level imbalance problem in typical surround sound material. A first AEC system, based on the MC-FDAF, that can efficiently cope with surround sound by taking into account all cross-correlations into the adaptation procedure has been presented in [4, 5]. In the following, both a suitable generalization of the above-introduced preprocessing approach is presented and the level imbalance problem is addressed. Section 4 shows that these elements may be efficiently integrated into a surround sound system.

3.2.1. Multi-Channel Preprocessing

The proposed perceptual phase modulation preprocessing technique can be adapted to multi-channel audio, such as the popular 5.1 surround format. Such setups generally include a Left front (L), Right front (R), Center (C), Left Surround (Ls), and Right Surround (Rs) speaker (plus a low frequency channel). To this end, phase modulation is carried out by using three independent modulators which modulate the L/R channel pair, the Ls/Rs channel pair and the C channel, respectively. Similarly to Figure 2, the modulation of channel pairs is carried out in a complex conjugate fashion. The modulation frequencies of the three modulators are chosen such that they are not commensurate with each other and thus provide "orthogonal" modulation activity. As an example, a modulation period of 1.3s was used for L/R processing, 3s for C processing and 1.1 s for Ls/Rs processing. This modular approach may also be extended to surround sound formats with more channels (e.g. 7.1) by adding further phase modification pairs after Fig. 2 with different modulation periods.

3.2.2. The Level Imbalance Problem

In surround sound, there is typically a significant and permanent level difference among the channels. In contrast to the front channels, the surround (back) channels often carry only weak ambience signals. Consequently, these channels are 'penalized' in their coefficient convergence. Although this problem is far more significant for surround sound than for stereo, it was described before for two-channel stereo, and some modifications of the simple normalized least-mean-squares (NLMS) algorithm have been proposed [17, 18].

For more flexibility w.r.t. the adaptation algorithm, we addressed the imbalance problem in our experiments by a channel-wise normalization according to a running recursive power estimate with a relatively large time constant, e.g., 40 s, so that only long-term changes are captured and the AEC convergence is not compromised. Our experimental studies based on the MC-FDAF [4, 5] have shown that this simple real-time capable solution delivers a significant performance improvement that is comparable to an individual offline normalization of each channel.

4. EXPERIMENTAL RESULTS

In order to illustrate the benefit of the proposed frequency selective phase modulation, we consider the performance in terms of both the convergence of the adaptive filter coefficients and the subjective quality of the preprocessed loudspeaker signals by means of a standardized listening test. To begin with, we compare the convergence of the coefficient misalignment

$$\epsilon(k) = \frac{\sum_{i=1}^{P} \|\mathbf{h}_i - \hat{\mathbf{h}}_i(k)\|^2}{\sum_{i=1}^{P} \|\mathbf{h}_i\|^2}$$

over time with different preprocessing methods. This comparison is carried out for both a stereo and a 5-channel AEC based on the MC-FDAF as described in detail in [5]. In all simulations, the echo-to-background noise ratio in the receiving room was set to 30 dB, and the regularization of the algorithm was optimized so that stability is provided for all preprocessing methods. In all our experiments, the sampling rate of the loudspeaker signals and the preprocessing was 44.1 kHz, while for the microphone channel and the echo cancellation it was downsampled by a factor of 4 as typical for speech recognition applications. The filter length L, covering the reverberation in the receiving room, was set to 1024.

Figure 4 shows the performance comparison in the stereo case for reproducing a quasi-monophonic high-quality male speech signal with an alternating spatial position in the transmission room (see Fig. 1). As baseline data, both a convergence curve without any preprocessing (curve label "without preproc.") and the well-known preprocessing by the nonlinearity after [7] (nonlinearity parameter $\alpha = 0.5$, label "NL") are included. Furthermore, a broadband phase modulation preprocessing (label "Pmod") and a perceptually tuned frequency selective phase modulation (label "Pmod_fs") are included, both with a sine modulation of 0.75 Hz. The modulation strengths of both systems have been adjusted in an informal listening procedure to deliver comparable spatial sound fidelity. As it can be seen from the data, convergence without any preprocessing is extremely slow, and using the non-linearity results in a significant convergence boost. Compared to this, the broadband phase modulation delivers a similar convergence enhancement within the first 15 s of the measurement. Finally, the perceptually tuned phase processing achieves the fastest convergence in this interval.

To compare the coefficient convergence for multi-channel application, we again chose the critical test signal of two alternating speaker positions in the transmission room. Note that in the surround sound scenario, it is important to choose a realistic recording scenario in order to take the level imbalance problem into account. Our setup in the transmission room was thus inspired by the so-called Decca Tree and surround microphones [19]. Figure 5 shows the corresponding coefficient convergence for various preprocessing methods. In addition to the methods considered in the stereo case,

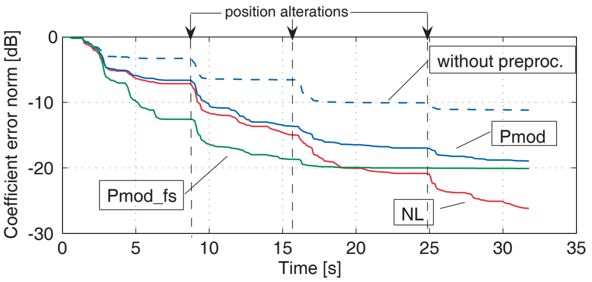

Fig. 4. Convergence comparison for stereo AEC.

we also consider the addition of uncorrelated noise after [9], labeled 'mp3_48' (individual mp3 en/decoding of each audio channel at 48 kbit/s). Again, the regularization was adjusted so that stability is provided for all preprocessing methods. We see that with the chosen set of parameters, all preprocessing methods considered here yield very similar convergence characteristics which is a good basis for our comparison in terms of a subjective listening test.

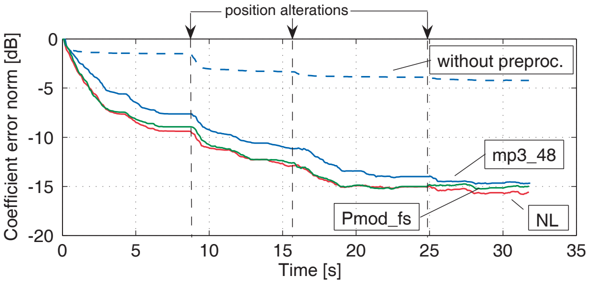

Fig. 5. Convergence comparison for 5-channel AEC.

Fig. 6 illustrates the results of the standardized subjective listening test called 'MUlti Stimulus test with Hidden Reference and Anchor' (MUSHRA) [20] assessing the sound quality of preprocessed 5-channel surround sound material for various preprocessing algorithms. The listening test was performed by 10 (9 of them experienced) listeners in a typical surround sound setup after [21]. The sound quality is quantified on a scale from 0 (very bad quality) to 100 (indistinguishable from original) for 5 critical music excerpts (items 'fount', 'glock', 'indie', 'pops', 'poule') which are also known in the MPEG context (see, e.g., [22]), one speech excerpt ('spmg') as explained above, and the average over all items. The different preprocessing types are the original reference and a 3.5 kHz band-limited version thereof (both included as required by [20]), individual channel mp3 en/decoding at 48 kbit/s ('mp3_48'), the novel perceptual phase modulation method ('phase'), a combination of mp3 encoding/decoding and phase modulation ('mp3_48_phase') and the conventional non-linear processing ('NL' after [7, 8]). It is visible from the graph that the phase modulation method emerges as the clear winner in terms of sound quality (note also the very low score for NL at the glockenspiel item due to objectionable artifacts). Furthermore, it can be combined with other preprocessing methods without noticeable further degradation of the sound quality.

5. CONCLUSIONS

A novel perceptually motivated preprocessing method for multi-channel acoustic echo cancellation has been presented. The experimental evaluation by a MUSHRA listening test shows superior perceptual quality without sacrificing convergence speed compared to state-of-the-art approaches. The proposed method also generalizes well to multi-channel, and thus surround sound systems.

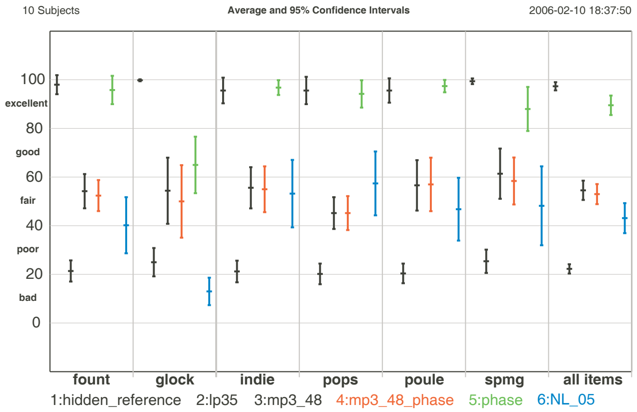

Fig. 6. Subjective audio quality for preprocessing methods.

6. REFERENCES

[1] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," *Proc. IEEE ICASSP,* pp. 3059-3062, May 1995.

[2] T. Gänsler and J. Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *Int. Journal of Adaptive Control and Signal Processing,* vol. 14, pp. 565-586, 2000.

[3] A. Sugiyama, Y. Joncour, and A. Hirano, "A stereo echo canceller with correct echo path identification based on an input-sliding technique," *IEEE Trans. Signal Processing,* 49(1), pp. 2577-2587, 2001.

[4] H. Buchner and W. Kellermann, "Acoustic echo cancellation for two and more reproduction channels," *Proc. Int. Workshop on Acoustic Echo and Noise Control,* pp. 99-102, Sept. 2001.

[5] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Processing,* vol. 85, no. 3, pp. 549-570, March 2005.

[6] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation- An overview of the fundamental problem," *IEEE Signal Processing Lett.,* vol. 2, pp. 148-151, Aug. 1995.

[7] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing,* vol. 6, pp. 156-165, Mar. 1998.

[8] D.R. Morgan, J.L. Hall, and J. Benesty, "Investigation of Several Types of Nonlinearities for Use in Stereo Acoustic Echo Cancellation," *IEEE Trans. Speech Audio Processing,* vol. 5, no. 6, pp. 686-696, Sept. 2001.

[9] T. Gänsler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," *Proc. IEEE ICASSP,* pp. 3649-3652, 1998.

[10] J. Benesty et al., "Stereophonic acoustic echo cancellation using nonlinear transformations and comb filtering," *Proc. IEEE ICASSP,* pp. 3673-3676, 1998.

[11] C. Neubauer, J. Herre, "Digital Watermarking and Its Influence on Audio Quality," *105th AES Convention,* San Francisco 1998, Preprint 4823.

[12] H.S. Malvar, "A modulated complex lapped transform and its application to audio processing," *Proc. IEEE ICASSP,* pp. 1421-1424, 1999.

[13] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," *Proc. IEEE ICASSP,* pp. 3689-3692, 1998.

[14] T. Hoya, J.A. Chambers, and P.A. Naylor, "Low complexity ϵ-NLMS algorithms and subband structures for stereophonic acoustic echo cancellation," *Conf. Rec. Int. Workshop on Acoustic Echo and Noise Control (IWAENC),* pp. 36-39, 1999.

[15] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization,* revised edition, MIT Press, 1997.

[16] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.,* vol. 11, no. 6, Nov. 2003.

[17] A. Nakagawa and Y. Haneda, "A study of an adaptive algorithm for stereo signals with a power difference," *Proc. IEEE ICASSP,* vol. 2, pp. 1913-1915, 2002.

[18] T. Fujii and S. Shimada, "Linear-combined multi-channel adaptive digital filters," *IEICE Trans.*, J69-A, no. 10, Oct. 1986.

[19] R. Streicher and F. Alton Everest, *The new stereo soundbook,* second ed., Audio Engineering Associates, Pasadena, CA, 1998.

[20] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," Jan. 2003.

[21] ITU-R, "Recommendation BS.1116-1: Method for the subjective assessment of small impairments in audio systems incl. multichannel sound systems," Oct. 1997.

[22] J. Breebaart et al., "MPEG spatial audio coding / MPEG Surround: overview and current status," *119th AES Convention,* New York 2005, Preprint 6599.