



APPLICATIONS OF DIRECTIONAL AUDIO CODING IN AUDIO

PACS: 43.60.Dh

Pulkki, Ville;

Lab. Acoustics and Audio Signal Processing, Helsinki University of Tech., POBox 3000, FI-02015, Finland

ABSTRACT

Directional Audio Coding (DirAC) is a method for spatial sound representation, applicable to different sound reproduction systems. In the analysis part, the diffuseness and direction of arrival of sound is estimated in a single location depending on time and frequency. In the synthesis part, microphone signals are first divided into non-diffuse and diffuse parts, and are then reproduced using different strategies. In this paper, the technology is reviewed, and implementations of DirAC for different applications are described.

INTRODUCTION

The coincident microphone approaches [1], such as first-order Ambisonics [2], can in theory utilize any loudspeaker setup. They are an attractive approach in spatial sound reproduction due to availability of good-quality microphones. Unfortunately, since the directional patterns of current high-quality microphones are only of zeroth or first order, the resulting loudspeaker signals of a multichannel setup are more coherent than desired. High coherence results in coloring and distortion of the spatial image, especially outside of the best listening position.

The goal of the proposed Directional Audio Coding (DirAC) is to reproduce the spatial properties of sound recorded with a B-format microphone system as realistically as possible with different sound reproduction systems. DirAC performs some analysis of the sound field at frequency bands, and uses that information in reproduction, which circumvents some problems in traditional techniques. DirAC is based on the same principles and partly the same methods as the recently-proposed Spatial Impulse Response Rendering (SIRR) technique [3]. SIRR is a technique to reproduce room impulse responses for application in convolving reverberators. The first attempts to reproduce other sound than impulse responses with SIRR were presented in [4], which later lead to a system which is currently called Directional Audio Coding.

Although DirAC was initially designed for high-quality applications, such as music reproduction etc., there are also some other applications where the technique can be used. As discussed in this paper, the technique can be used to transmit the spatial aspects of audio analyzed from a low-cost microphone grid as metadata, while the transmitted audio would only be a monophonic channel. This enables teleconferencing applications with spatial sound, as the transmitted data rate would not increase prominently from current solutions using also monophonic audio transmission.

ASSUMPTIONS ABOUT PERCEPTION OF SPATIAL SOUND

The design of the SIRR and DirAC techniques is based on four assumptions about the interaction between sound field properties and perceptual attributes that they produce [3]. The assumptions are repeated briefly here. The interested reader is referred to [5] for a discussion about their validity and for an introduction to the human mechanisms for spatial sound perception relevant in this case. For a more concise understanding of the perceptual mechanisms the reader is referred to, e.g. [6, 7].¹ The assumptions are: 1 Direction of arrival of sound transforms into interaural time difference (ITD), interaural level difference (ILD), and monaural localization cues. 2 Diffuseness of sound transforms into interaural coherence cues. 3 Timbre depends on the monaural (time-dependent) spectrum together with ITD, ILD, and interaural coherence of sound. 4 The direction of arrival, diffuseness, and spectrum of sound measured in one position with the temporal and spectral resolution of human hearing determines the auditory spatial image the listener perceives.

In SIRR, the number of sources used to measure the impulse response was always one, and the signal

¹It has to be noted, that the term "diffuse" denotes only a quality of physical sound field in this paper. It is not used to quantify any perceptual aspect.

content of it was known to be an impulse. In DirAC, multiple sources are also allowed, and their content is not a priori known. However, the assumptions are not changed. It is still assumed that the listener cannot decode separate cues for multiple wavefronts from different directions within a critical band at one time instant. This is bolstered by psychoacoustic results, where it has been found, that two sinusoids with small spectral difference presented over spatially separated loudspeakers can not be localized correctly [8]. The listener typically localizes the sinusoids as a single fused event fluctuating inside the head between the ears. This makes us to assume that humans decode at one time only single cues per each critical band from ear canal signals.

DirAC PROCESSING

In this section, the processing parts of DirAC are presented. The general idea is first reviewed, after which different parts of processing are considered separately. Since there are multiple applications for DirAC, there are multiple variations for some of the parts.

General idea of DirAC

The DirAC design is based on the assumptions stated earlier in this paper, as was the SIRR design. In assumption 4 it was stated that the temporal and spectral resolution of the processing should mimic the temporal and spectral resolution the auditory system is using for spatial hearing. For this purpose, the microphone signals are divided into frequency bands, following the frequency decomposition of the inner ear as shown in Fig. 1.

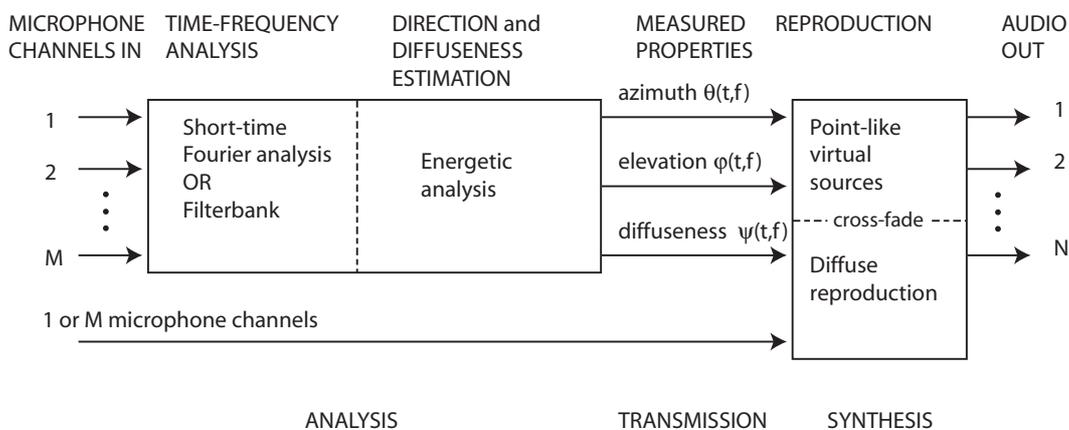


Figure 1: Overall flow diagram of DirAC.

The assumptions also imply that the direction of arrival, diffuseness, and spectrum of the sound field should be synthesized correctly, which would lead to generation of correct spatial cues to the listener. Thus, the direction of arrival and diffuseness are analyzed with a temporal accuracy comparable to the accuracy of the human auditory system at each frequency band. This information is later used in DirAC synthesis to generate these cues correctly. In practice, in synthesis the sound is dynamically divided into two streams, one stream which corresponds to non-diffuse sound, and another stream which corresponds to diffuse sound stream. The non-diffuse sound stream is reproduced with a technique aiming at point-like sound sources, and the diffuse sound stream with a technique aiming at perception of sound lacking prominent direction.

Between analysis and synthesis, the sound may be transmitted over any medium. This possibility is exploited in teleconferencing application, where the audio is transmitted as monophonic channel with analysis parameters as metadata.

The microphone array used with DirAC must enable the analysis of direction and diffuseness in a broad frequency region. So far B-format microphones have been solely used. B-format stands for a coincident microphone array which produces four microphone channels with different directional characteristics: one omnidirectional, and three figure-of-eight channels directed towards each orthogonal axis of Cartesian coordinates. The corresponding signals are $w(n)$, $x(n)$, $y(n)$ and $z(n)$, respectively, where n is the time index. The omnidirectional signal has been scaled by $1/\sqrt{2}$. In time-frequency analysis the microphone channels are divided into frequency bands, which are denoted, e.g., by $w(n, i)$, where i denotes the frequency channel index. However, for simplicity of notation, the indices are dropped where appropriate.

With B-format input, the directional analysis can be performed relatively simply based on energetic analysis of sound field, which is now shortly reviewed theoretically, and then with B-format input. Instantaneous

energy density can be computed as $E = \frac{1}{2}\rho_0 \left(\frac{p^2}{Z_0^2} + u^2 \right)$, where p is sound pressure, u is particle velocity ρ_0 is the mean density of air, and Z_0 is the characteristic acoustic impedance of air $Z_0 = \rho_0 c$, where c is the speed of sound [9]. The instantaneous intensity vector \mathbf{I} is defined as $\mathbf{I} = p\mathbf{u}$, where \mathbf{u} is particle velocity vector [9]. The intensity vector points to the direction of net flow of energy, and the magnitude denotes the strength of energy flow. Using these equations, it is possible to define diffuseness

$$\psi = 1 - \frac{\|\langle \mathbf{I} \rangle\|}{c\langle E \rangle}, \quad (1)$$

where $\langle \cdot \rangle$ denotes short-time average. Diffuseness gets a value of zero with plane waves from a single direction, where the net flow of energy corresponds to total energy. It reaches the value of one in a field, where there is no net transport of acoustic energy, as with ideal standing waves or reverberation.

With B-format input, p is approximated by w , and particle velocity vector by $\mathbf{v} = x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z$, where \mathbf{e}_x , \mathbf{e}_y and \mathbf{e}_z represent Cartesian unit vectors. The diffuseness is computed from B-format input with

$$\psi(n) = 1 - \frac{\sqrt{2} \sum_{m=a_1}^{b_1} \|w(n+m)\mathbf{v}(n+m)\| W_1(m)}{\sum_{m=a_1}^{b_1} (|w(n+m)|^2 + |\mathbf{v}(n+m)|^2) W_1(m)/2}. \quad (2)$$

where $W_1(m)$ is a window function defined between constant time values $a_1 \leq 0$ and $b_1 > 0$ for short-time averaging explained in the next section. The direction vector is defined to be the opposite direction of intensity vector. In this case, the length of the intensity vector is not of interest which simplifies the formula a bit. The vector is computed at each frequency channel as

$$\mathbf{D}(n) = - \sum_{m=a_2}^{b_2} w(n+m, m)\mathbf{v}(n+m)W_2(m), \quad (3)$$

where W_2 is a window function for short-time averaging \mathbf{D} , and a_2 and b_2 defined similarly as a_1 and b_1 , respectively. For a more detailed presentation of energetic analysis with B-format input, see [10].

In SIRR, the temporal variations of the analyzed direction and diffuseness were not limited. If such variations were applied in DirAC for reproduction of continuous sound, there would be audible artifacts. In Eqs. 2 and 3 a window function was used for short-time averaging. With DirAC implementations using STFT time-frequency analysis, the window has been implemented with a first-order low-pass IIR. With off-line processing, the window function can also be defined to be symmetric in time, as in the current high-quality implementation the Hanning window has been used. In this case, different lengths for window are applied at different frequency bands. The variables from which the diffuseness is computed are temporally averaged, as shown in Eq. (2).

The slowing down of direction vector \mathbf{D} is a slightly more complicated task. The reproduction of direction has to be accurate. In particular, during the onsets of continuous sounds, analyzed direction has to correspond accurately to the direction of arrival of sound. However, the changes in direction may be rapid in many cases. If such fast changes are applied directly in synthesis, temporal artifacts, such as clicks and bubbling emerge.

If the direction vector \mathbf{D} itself is temporally averaged, fast changes in analyzed direction are replaced by slow changes. The vector will then most of the time point to a direction which does not correspond to the instantaneous direction-of-arrival of sound. In informal tests, averaging of \mathbf{D} was found suboptimal, as the temporal artifacts were not completely removed, and the perceived sound source directions were instable. A solution to this problem is to apply for temporal averaging of the direction vector with as short time window as is feasible to compute the direction of arrival reliably. In addition, to avoid audible clicks and other artifacts, the reproduction method parameters are averaged with a longer time window. In the current implementation, the gain factors of amplitude panning are averaged over time.

Synthesis techniques

In DirAC synthesis, the microphone signal(s) are divided into non-diffuse and diffuse streams by multiplying the input signal(s) with two time-variant factors derived from diffuseness. The diffuse stream will be reproduced in random phase when compared to non-diffuse stream, and the cross-fading between the streams has to conserve energy. This can be simply implemented by multiplying input signal(s) with $\sqrt{1-\psi}$ and $\sqrt{\psi}$, respectively, as shown in Figs. 2 and 3. In practice, the analyzed diffuseness has never the value of

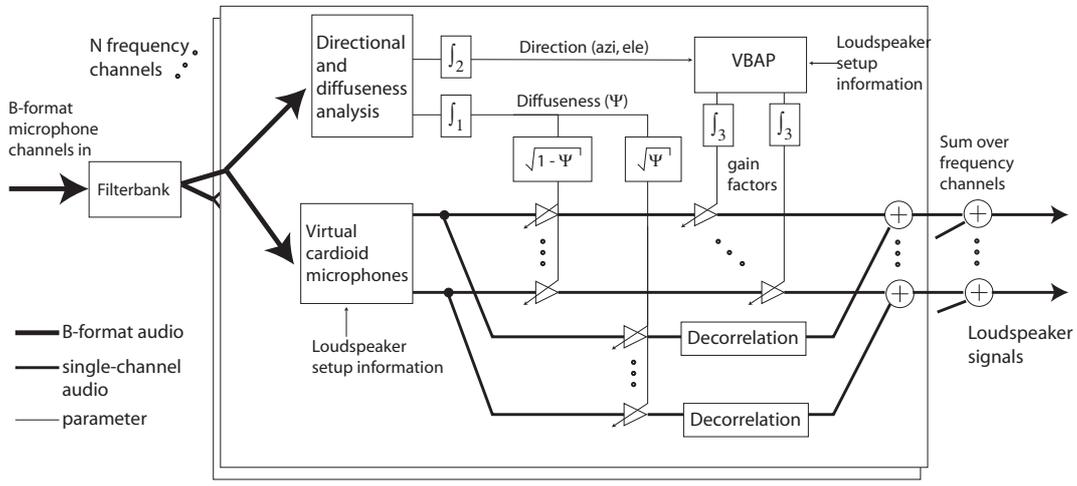


Figure 2: Flow diagram of DirAC for high-quality reproduction of B-format sound. The integral symbols with a subscript denote temporal averaging with corresponding window function.

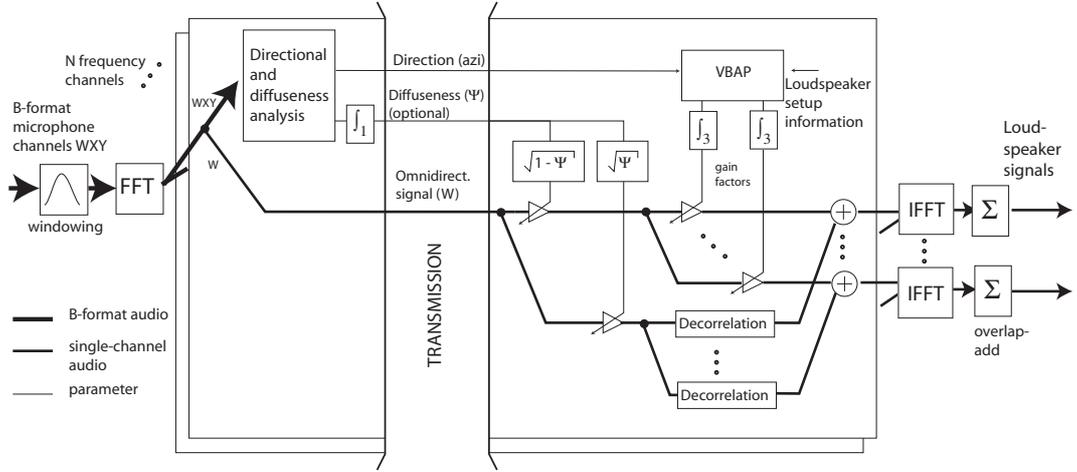


Figure 3: Flow diagram of DirAC for telecommunication. The integral symbols with a subscript denote temporal averaging with corresponding window function.

one or zero, but lies in between. Thus, the produced loudspeaker signals cross-fade at each frequency band between signals reproduced with diffuse and non-diffuse methods.

Reproduction of point-like virtual sources

The choice of the method to create point-like sources is dependent on targeted sound reproduction equipment. In this work, relatively sparse loudspeaker 2-D and 3-D systems have been available, thus amplitude panning formulated with Vector Base Amplitude Panning (VBAP) [11] has been considered sufficient. With amplitude panning, the computation is straightforward, the 2-D or 3-D directional information is given to VBAP algorithm, which computes the gain factors $g(n, i, k)$ for each time instant n , frequency channel i and loudspeaker channel k as schematically shown in Fig. 2 or in Fig. 3.

The second phase of temporal averaging in directional reproduction has to be computed in the synthesis phase. The averaging is now described for amplitude panning. The artifacts are avoided by computing weighted temporal-domain average of the gain factors from VBAP:

$$g(n, i, k) = \frac{\sum_{m=-M/2}^{M/2} g(n+m, i, k)(1-\psi(n+m, i))W_3(m)}{\sum_{m=a_3}^{b_3} (1-\psi(n+m, i))W_3(m)}, \quad (4)$$

where $g(n, i, k)$ is the gain factor at subband i , time n and loudspeaker channel k . With filterbank time-frequency analysis, the length of Hanning window $W_3(m)$ has been as high as 100 times the period of center

frequency of the corresponding channel, but limited to 1000 ms at maximum. However, the weighting with $(1 - \psi)$ effectively shortens the time window when an onset with low diffuseness occurs. The weighting factor in Eq. (4) can also be, e.g., instantaneous energy E instead of $(1 - \psi)$. If the diffuseness is not transmitted in the teleconferencing application, instantaneous energy is the only choice in that case.

In the teleconferencing implementation, the only channel which is transmitted is w , and it is the audio signal for point-like virtual sources after optional multiplication by $\sqrt{1 - \psi}$, as shown in Fig. 3. This has one shortcoming: only in the case when diffuseness reaches the value of zero, which means that in recording phase there was sound from only one direction in this time instant, w includes only non-diffuse sound. When diffuseness is greater than zero, which is generally the case, w is a superposition of diffuse and non-diffuse sound. Thus the sound which is applied to point-like virtual source includes also some diffuse sound, which is not desired.

This defect can be partly avoided when more channels are transmitted. In the high-quality implementation B-format signals are transmitted, thus the audio signal for a point-like virtual source can be derived from a virtual cardioid microphone directed towards the analyzed direction. This will reduce the amount of diffuse sound energy in average by 4.8 dB [12]. In practice, this is implemented by computing a signal for each loudspeaker $y(n, i, k)$ with equation

$$y(n, i, k) = \frac{1}{2}g(n, i, k)[w(n, i)\sqrt{2} + x(n, i)\cos\theta_k\cos\phi_k + y(n, i)\sin\theta_k\cos\phi_k + z(n, i)\sin\phi_k], \quad (5)$$

where θ_k is the azimuth and ϕ_k is the elevation of loudspeaker k . A system implementing this is illustrated in Fig. 2

The aim of diffuse synthesis is to produce perception of surrounding sound lacking prominent direction. Also, the processing should not introduce any coloration to the sound spectrum. With multichannel listening, such perception can be produced by applying a sound signal that is decorrelated to all loudspeakers. In that case, the magnitude spectrum of sound should not be changed, but the phase spectrum should be random.

In this project a simple, although computationally quite complex method for decorrelation is to convolve sound for each loudspeaker with a random sequence, which was also one of methods used with SIRR. The sequences must be different for each loudspeaker, and they should be composed in a way that the convolution process would change the phase at all frequencies. These prerequisites are achieved by having short decay times for high frequencies in the sequence, and longer decay times for low frequencies.

When only a single channel is transmitted, as in the teleconferencing implementation of DirAC, the diffuse stream may be composed only from the single channel multiplied by $\sqrt{\psi}$. In the current implementation for multichannel loudspeaker setups, the audio channel is simply decorrelated for each loudspeaker, as shown in Fig. 3. Using only one signal as a source for diffuse sound may be suboptimal. The diffuse sound may have a directional distribution, which should be reproduced, and in some cases the sound arriving from different directions has different content.

If B-format signal is transmitted, as in high-quality implementation of DirAC, these factors are reproduced more faithfully to the original case. A virtual cardioid microphone is computed pointing towards each loudspeaker direction and the signals are applied after decorrelation to corresponding loudspeakers. The decorrelation is needed to decrease the coherence between loudspeaker signals which is due to the broadness of applied first-order microphone directional patterns.

As different reproduction methods are applied to non-diffuse and diffuse sound, their outputs may have to be scaled with a constant factor to maintain an even balance between the methods. This scaling is not presented in the figures, since it can be performed by adjusting the reproduction methods, e.g. by scaling the gain factors of amplitude panning, or by scaling the random sequences used in decorrelation. In the current implementation, a factor $1/\sqrt{N}$, where N is the number of loudspeakers, is used as the factor for random sequences used in decorrelation.

APPLICATIONS

A natural application for DirAC is the delivery of spatial sound for arbitrary domestic listening conditions. In the production of audio material, DirAC provides new types of processing of B-format recordings. The spatial image can be modified after directional analysis by modifying the analyzed directions, or amplifying sound depending on direction. Also, as the diffuse and non-diffuse streams are separated, they can be weighted differently to emphasize either dry or reverberant sound. This also enables different effects on different streams, e.g., diffuse sound, which is mostly reverberant sound, can be equalized differently than non-diffuse sound.

In teleconferencing, the application of DirAC is interesting, when there are multiple persons at least in one end of the transmission line. The aim would be that the directions of the talkers are reproduced to increase the intelligibility in multi-talker or noisy environments, and to increase the naturalness of connection. When three or more parties are connected together, the joining of the streams would be necessary. In joining the intensity vectors and energies of DirAC streams have be computed from transmitted data, which can then be added together, and new metadata can be computed as in directional analysis.

The DirAC processing makes it also possible to amplify the sound depending on the analysed direction. This can be interpreted as a new type of directive microphone. In informal tests, where three talkers were recorded with a simulated microphone in directions -30° , 0° and 30° , the individual talkers could be amplified by 10-20 dB without audible artifacts.

In principle, DirAC can also be used to perform parametric multichannel surround coding to encode multi-channel audio files into a stream with one or three audio tracks similarly as in [13]. The encoding can be computed in similar fashion to stereo upmixing, the listening setup is simulated in anechoic B-format recording, and DirAC is applied to estimate directions and diffusenesses at frequency bands. However, this has not been tested.

Conclusions

The Directional Audio Coding (DirAC) technique is presented in this paper. DirAC is a technique to reproduce spatial sound over different reproduction systems, and it can also be used to encode spatial aspects of sound as metadata which is transmitted/stored along with a single or more audio channels. The technique is based on analyzing the sound direction and diffuseness depending on time at narrow frequency bands, and further using these parameters in sound reproduction with appropriate techniques.

Acknowledgments

Ville Pulkki has received funding from the Academy of Finland (project 105780) and from Emil Aaltonen foundation.

References

- [1] S. P. Lipshitz. Stereo microphone techniques... Are the purists wrong? *J. Audio Eng. Soc.*, 34(9):716–744, 1986.
- [2] M. A. Gerzon. Periphony: With-Height Sound Reproduction. *J. Audio Eng. Soc.*, 21(1):2–10, 1973.
- [3] J. Merimaa and V. Pulkki. Spatial Impulse Response Rendering I: Analysis and synthesis. *J. Audio Eng. Soc.*, 53(12):1115–1127, 2005.
- [4] V. Pulkki and J. Merimaa. Spatial impulse response rendering: Listening tests and applications to continuous sound. In *AES 118th Convention*, Barcelona, Spain, 2005. Preprint 6371.
- [5] V. Pulkki and J. Merimaa. Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests. *J. Audio Eng. Soc.*, 54(1/2):3–20, 2006.
- [6] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.
- [7] D. W. Grantham. Spatial hearing and related phenomena. In B. C. J. Moore, editor, *Hearing*, pages 297–345. Academic press, 1995.
- [8] D. R. Perrott. Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays. *J. Acoust. Soc. Am.*, 76(6):1704–1712, Dec. 1984.
- [9] F. J. Fahy. *Sound Intensity*. Elsevier Science Publishers Ltd., Essex, England, 1989.
- [10] J. Merimaa. *Analysis, synthesis, and perception of spatial sound — binaural localization modeling and multichannel loudspeaker reproduction*. PhD thesis, Helsinki Univ. Tech., 2006. Available at: <http://lib.tkk.fi/Diss/2006/isbn9512282917/>.
- [11] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6):456–466, 1997.
- [12] R. B. Schulein. Microphone considerations in feedback-prone environments. *J. Audio Eng. Soc.*, 24(6):434–445, July/August 1976.
- [13] M. Goodwin and J. Jot. A frequency-domain framework for spatial audio coding based on universal spatial cues. In *AES 120th Convention*, Paris, France, 2006. Convention paper # 6751.