A COMPUTATIONAL MODEL FOR THE ESTIMATION OF LOCALISATION UNCERTAINTY

Enzo De Sena, Zoran Cvetković

Institute of Telecommunications, King's College London, London, WC2R 2LS, UK
enzo.desena@kcl.ac.uk, zoran.cvetkovic@kcl.ac.uk

ABSTRACT

A computational model for prediction of localisation uncertainty of phantom auditory sources is proposed. The interaural level and time difference pairs due to point sources in free field are used as a reference. The mismatch between these "natural" pairs and interaural time and level difference pairs elicited by phantom sources is quantified by means of the 0.5-norm distance, which is justified on psychoacoustic grounds. The model is validated by results of subjective listening tests, achieving a high level of correlation with experimental data.

Index Terms— Auditory model, perception, simulation, naturalness, locatedness, localisation, multichannel audio.

1. INTRODUCTION AND RELATED WORK

Subjective listening experiments are the most reliable method for studying perceptual auditory phenomena, but they require very careful design and carrying them out is expensive and time-consuming [1]. Computational models provide a fast and repeatable alternative. However finding reliable models is very challenging due to the complexity of the human auditory system.

Computational models are not only useful for predicting the outcome of otherwise cumbersome listening experiments. Finding models that correctly replicate experimental results can, to some extent, shed some light into the principle of sound processing in the auditory system. Towards understanding spatial hearing, a number of models have been developed over the past sixty years. The first notable attempt in this direction was made by Jeffres in 1948 [2]. He hypothesised that sound source localisation is governed by a mechanism of running cross-correlation between the two ear signals. While Jeffres' model is still considered to be an adequate way of measuring interaural differences in wavefront arrival times (ITD), it does not account for the other important cue in source localisation, *i.e.* interaural level differences (ILD) [3]. In the 1980s Lindemann proposed a model that incorporates this information in the cross-correlation mechanism by means of physiologically plausible inhibitory elements [4].

Gaik [5] extended this model further based on the observation that interaural level and time differences due to natural sources (*i.e.* point-like sources in free field) come in specific pairs. For instance the ILD and ITD values for a source on the median plane are both small. On the other hand, for a source to the right/left, both ILD and ITD are high; in these cases the sound wave arriving at the far ear is both attenuated (because of head shadowing) and delayed (because of propagation time) compared to the sound wave arriving to the ear closer to the sound source.

Unnatural ILD-ITD combinations can be delivered artificially through headphones. Gaik observed that in these cases the width of the auditory event increases and sometimes two separate events are reported [5, 3]. These unnatural conditions can arise also in cases other than headphones presentation, creating the impression of the diffuseness of the sound source and uncertainty about its location. This is the case for most sparse multichannel reproduction systems where a small number of loudspeakers (the simplest case being stereophony) are used to give the impression of an acoustic source located somewhere between them [3]. Quantifying the mismatch between the reproduced ILD-ITD pairs and the ones associated to natural sources is therefore of particular interest for the design of multichannel technologies.

A study presented by Pulkki and Hirvonen in [6] goes in this direction. For a given multichannel system they find the angle of the closest natural source in terms of ILD and ITD, separately. This model can give some useful predictions when these two angles coincide. However, in most cases the ILD and ITD cues provide contradicting information, and therefore the model output is hard to interpret [6]. As in [6] we also assume that the auditory system uses the natural ILD-ITD pairs as reference points. However we propose a novel way of using the ILD and ITD cues jointly. In order to calculate the discrepancy between the natural ILD-ITD pairs and the measured ones, we use the 0.5-norm distance, and we justify this choice on psychoacoustic grounds. On the basis of this distance, the model yields predictions of localisation uncertainty.

The paper is organised as follows. In Section 2 the proposed computational model is described. The model is then validated in Section 3 by means of a comparison with experimental data. Conclusions are drawn in Section 4.

2. COMPUTATIONAL MODEL

The proposed computational model measures the localisation uncertainty of phantom images by means of a distance between the ILD-ITD pairs in critical bands of hearing associated to natural sources and ILD-ITD pairs elicited by systems under investigation. To this end, first a methodology for calculating ILD and ITD values in critical bands is developed. The same methodology will be used to calculate the ILD-ITD pairs due to natural sources as well as the ILD-ITD pairs due to signals produced my multichannel audio systems. Once a reference database of natural ILD-ITD pairs is built, a distance measure which quantifes the discrepancy between natural and stereophonic pairs in individual critical bands needs to be defined. Finally, these distance functionals are aggregated across critical bands to obtain the overall prediction of localisation uncertainty.

The work reported in this paper was funded by EPSRC Research Grant EP/F001142/1. The authors would like to thank Yaqub Alwan for the useful discussion on the topic.

2.1. Calculation of ILD and ITD pairs

The ILD and ITD values are calculated as follows. Acoustic sources are modelled as point sources in the free field. The source-to-ear transfer functions are obtained using the spherical head model proposed by Duda and Martens [7]. This model provides a reasonable approximation of the diffraction around the head and has the advantage of providing the response for sources in any given direction. Databases of head-related transfer functions (HRTF), on the other hand, provide data only in a discrete number of directions [8], usually with a resolution of no more than 5 degrees.

The frequency range where the model operates is chosen according to the following considerations. The experimental evidence suggests that humans use two main mechanism for source localisation that are to a certain degree independent from one another [3, p.173]. The first interprets the interaural time shifts between the signals' fine structure and uses signal components below 1.6 kHz. The second interprets the interaural level differences and time shifts of the envelopes *jointly*. Between these two mechanism, the latter seems to be the dominant one for signals with significant frequency content above 1.6 kHz. Based on these considerations, we assume that the most significant contribution to localisation uncertainty comes from inconsistencies between the interaural cues above 1.6 kHz.

The response of the cochlea is modelled using a gammatone filter-bank [9] with 24 centre frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale between 1.6 kHz and 15 kHz [10]. Each bandpass signal is processed using Bernstein's model of neural transduction [11]. The resulting signals are then fed to 24 binaural processors that calculate the ILD and ITD values. The ILD is calculated as the energy ratio of the left and right channels [12]. The ITD is selected as the location of the maximum of the cross-correlation function evaluated over time lags between $[-0.7, 0.7]$ ms [5, 12].

All in all, the model produces a set of 24 ILD-ITD pairs for a given input. In the next section we examine the position of these pairs when the input is due to a natural source.

2.2. ILD-ITD pairs for natural sources

In Figure 1 we show the ILD-ITD pairs corresponding to 181 natural sources uniformly spaced between $[-\pi/2, \pi/2]$ (relative to the look direction) in the horizontal plane. The stimuli used to obtain these pairs are pink noise samples that are 500 ms long. Two main observations can be made. First, as described qualitatively in the introduction, the interaural cues are highly correlated. This is due to the concurrent effect of propagation and diffraction around the head [5]. Second, for sources outside the median plane, the ILD values increase with frequency. This is due to the increasing shadowing of the head with decreasing wave lengths [5].

Notice that the curves in Figure 1 appear compressed along the ordinate axis as compared to the ones reported by Gaik in [5]. This is due to the fact that Gaik calculated the ILD directly on the bandpass signals after the gammatone filtering. As in [12] we calculate the ILD on the signals after the neural transduction step, which is more plausible physiologically.

As in [6] and [12] we hypothesise that the auditory perspective is formed using the natural ILD-ITD pairs as reference points. The question that arises then is how to quantify the mismatch between given ILD-ITD pairs and the natural ones, or, in other words, how to calculate the distance between them. This is the subject of the next subsection.

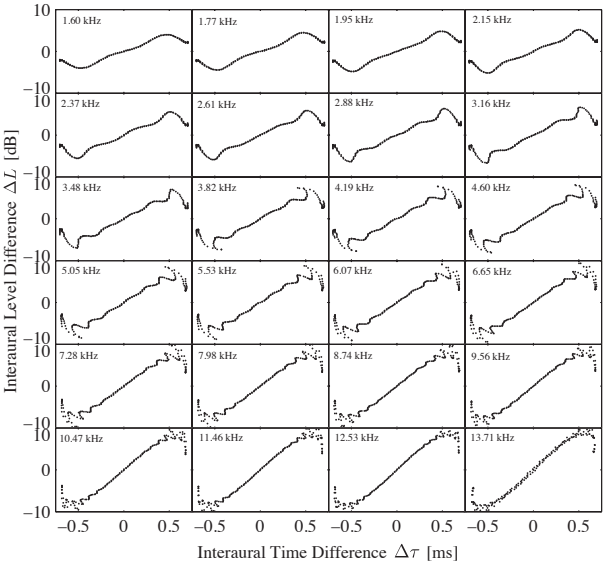


Fig. 1. The figure shows the ILD-ITD pairs associated to point-like sources in free field within each critical band. Each point corresponds to one of 181 natural sources uniformly spaced between $[-\pi/2, \pi/2]$ (relative to the look direction) in the horizontal plane.

2.3. Distance between ILD-ITD pairs

Let us denote the ILDs and ITDs in individual critical bands by ΔL_i and $\Delta\tau_i$, respectively, where i is the index of the critical band. The ILD and ITD associated to a natural source in the horizontal plane will be denoted as $\Delta L_i(\theta)$ and $\Delta\tau_i(\theta)$, respectively, where θ is the angle formed between the natural source and the listener's look direction.

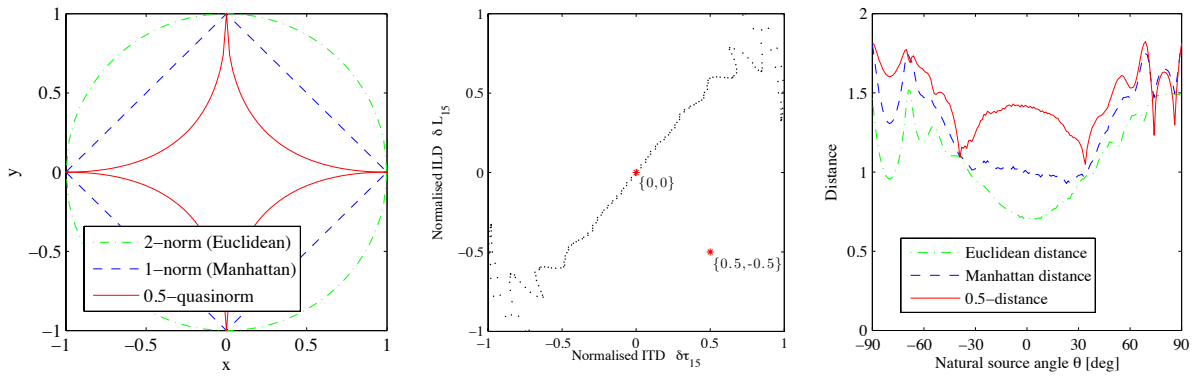
In order to combine the information of ILD and ITD cues across different bands, we first normalise ΔL_i and $\Delta\tau_i$ to the maximum natural values in the corresponding critical band, to obtain

$$\delta L_i = \frac{\Delta L_i}{\max_\theta |\Delta L_i(\theta)|}, \quad \delta\tau_i = \frac{\Delta\tau_i}{\max_\theta |\Delta\tau_i(\theta)|}\,. \qquad (1)$$

The natural values $\Delta L_i(\theta)$ and $\Delta\tau_i(\theta)$ are normalised in the same way and are denoted as $\delta L_i(\theta)$ and $\delta\tau_i(\theta)$, respectively. Note that while $\delta L_i(\theta) \in [-1, 1]$ and $\delta\tau_i(\theta) \in [-1, 1]$, generic δL_i and $\delta\tau_i$ can take values outside the range $[-1, 1]$.

We need now to find a distance functional between generic $\{\delta L_i, \delta\tau_i\}$ and natural $\{\delta L_i(\theta), \delta\tau_i(\theta)\}$ pairs according to some meaningful psychoacoustic criterion. There is experimental evidence that unnatural ILD-ITD combinations often trigger split auditory events [5]. More specifically, trained subjects report that one event (commonly referred to as "time image") is closely coupled with the ITD for its direction, while the other ("intensity image") depends on both ILD and ITD [3, p.170].

Consider the distance defined by the p-norm $||\mathbf{x}||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$. Unit spheres according to different p norms in $\mathbb{R}^2$ are illustrated in Figure 2a for $p = 2, 1, 0.5$. For $p = 2$, an unnatural point $\{0.5, -0.5\}$ results in the closest natural pair being $\{0, 0\}$ (see Figure 2b). The 1-norm has the potential to find a closest pair further away from the centre, however, the 0.5-norm distance results in two sharp minima, one of which is centred in the direction corresponding to the ITD cue, compatible with the psychoacoustic evidence



(a) (b) (c)

Fig. 2. Figure 2a shows the set of points (x, y) with unit distance from the centre for different p-norms, with $p = 2, 1, 0.5$. The normalised ILD and ITD are shown in Figure 2b for the 15th critical band. Figure 2c shows the distance between the point $\{\delta L_{15}, \delta\tau_{15}\} = \{0.5, -0.5\}$ and the natural points $\{\delta L_{15}(\theta), \delta\tau_{15}(\theta)\}$ as a function of θ for different distance functions.

(see Figure 2c). Other values of p close to 0.5 would also be acceptable. In the next section we show that, even without careful tuning, the model is capable of predicting experimental data accurately.

Notice that the p-norm $||\mathbf{x}||_p$ and the associated distance $||\mathbf{x} - \mathbf{y}||_p$ do not satisfy the triangular inequality for $p < 1$. It can be proven, however, that $||\mathbf{x} - \mathbf{y}||_p^p$ satisfies all the properties of a distance metric [13, p.301]. Therefore, for each of the critical bands we create the distance function $d_i(\theta)$, given by

$$d_i(\theta) = |\delta L_i - \delta L_i(\theta)|^{0.5} + |\delta\tau_i - \delta\tau_i(\theta)|^{0.5} \ , \qquad (2)$$

to quantify the distance between a considered pair $\{\delta L_i, \delta\tau_i\}$ and pairs $\{\delta L_i(\theta), \delta\tau_i(\theta)\}$ corresponding to natural sources at angles θ.

2.4. Quantifying the localisation uncertainty

In order to quantify the localisation uncertainty we proceed as follows. First, for each critical band a score function $\Gamma_i(\theta)$ of perceiving a sound source in direction θ is formed by flipping the corresponding distance function as

$$\Gamma_i(\theta) = \max_\theta \left(d_i(\theta)\right) - d_i(\theta) \ . \qquad (3)$$

Next, we need to combine the information associated with different critical bands. The mechanisms governing this stage of perception are generally regarded as very complex, and are not well understood [6, 3]. Hence we conservatively choose the overall score function as the simple average of the individual $\Gamma_i(\theta)$:

$$\Gamma(\theta) = \frac{1}{N} \sum_{i=0}^{N-1} \Gamma_i(\theta) \ , \qquad (4)$$

where N is the number of critical bands. In practical cases the ILD-ITD pairs of natural sources will be known for a discrete number of directions, which we denote as θ_k. The aggregated score function is then normalised according to

$$\gamma(\theta_k) = \frac{\Gamma(\theta_k)}{\sum_m \Gamma(\theta_m)} \qquad (5)$$

so that its sum across all angles is equal to one, $\sum_k \gamma(\theta_k) = 1$. This normalised function can be interpreted as proportional to the likelihood that an auditory event will be perceived in direction θ. From this perspective, a uniform $\gamma(\theta)$ would result in a maximally uncertain (diffuse) event. Based on this information-theoretic analogy, we define the localisation uncertainty as the entropy of $\gamma(\theta)$:

$$H = -\sum_k \gamma(\theta_k) \log_e \gamma(\theta_k) \qquad (6)$$

Notice that H could have been defined in various other ways. For instance, the score function (4) could have been chosen as $1/\left(\sum_i d_i(\theta)\right)$. In this case, however, the score function due to a natural source would be (with abuse of notation) $\delta(\theta)$. The localisation uncertainty H would in turn be equal to zero regardless of the direction θ, which is not in agreement with the experimental evidence showing that natural sources on the side result in higher localisation blurs [3]. On the other hand, the steps (3)-(4) correctly lead to the property that for natural sources H, and therefore localisation uncertainty, depends on θ. An alternative choice for H could be to define it as the average of the entropy calculated on the individual score functions of each critical band. However, that would not account for inconsistencies between critical bands, which are likely to increase localisation uncertainty.

3. COMPARISON WITH EXPERIMENTAL DATA

3.1. Locatedness experiment

In [14] a subjective listening test with 19 subjects was carried out. The test methodology was similar to that of MUSHRA [15] tests were the different stimuli are compared and graded at the same time. The perceptual attribute studied in this experiment was the source locatedness. Blauert defines locatedness as the "degree to which an auditory event can be said to be clearly perceived in a given direction" [3]. The subjects answered the question "How certain are you of the direction of the source" by giving a score on a continuous scale from 0 to 100. The scale was divided in five equal intervals labeled as "I am certain", "I have a slight doubt", "I have a doubt", "I am really not sure" and "I have no idea" [16].

The test was carried out in an audio booth using the setup shown in Figure 3. The experiment compared four synthesised multichannel technologies:

TPL Tangent panning law [17].

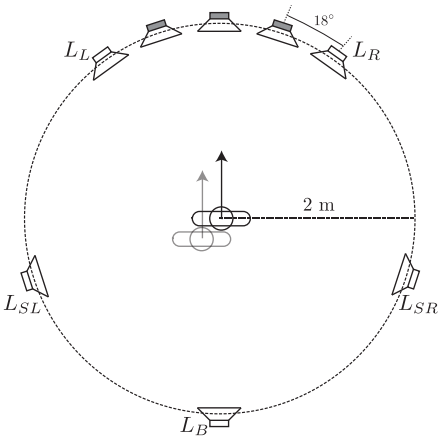


Fig. 3. The test setup for the locatedness experiment. The five uniformly spaced white loudspeakers form the reproduction system. The three dark loudspeakers in front of the listener are the acoustic pointers. Two listening position are considered. One in the centre looking in the midline direction between L_L and L_R. The second is 30 cm off-centre, and more specifically $30/\sqrt{2}$ cm behind and to the left of the central position. Adapted from [14].

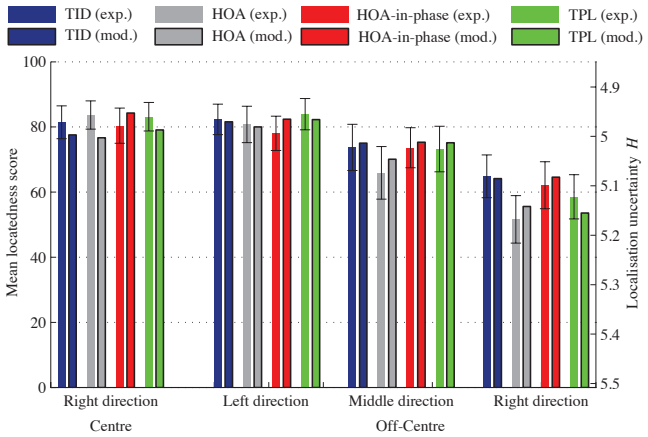


Fig. 4. Comparison of the listening experiment results from [14] with the localisation uncertainty predicted by the model. Notice the two different axes. The left axis is relative to the experimental data, while the right one is relative to the model predictions. The two axes were aligned and scaled using the linear regression of the two data sets (Pearson correlation is -0.94). The error bars represent the 95% confidence intervals of the mean locatedness score.

HOA Near-field corrected second-order Ambisonics with mode-matching decoding at low frequency and maximum-energy decoding at high frequency (cut-off frequency of 1200 Hz) [18, 19, 20].

HOA–in-phase Second-order Ambisonics with in-phase decoding [21].

TID The quasi-coincident microphone array proposed in [14], based on fitting of time-intensity psychoacoustic curves.

In order to assess the systems' performance under different conditions, the test was repeated for three source directions and two seating positions as shown in Figure 3. The three source directions corresponded to the direction of the acoustic pointers shown in Figure 3, labelled in the following as "left", "middle" and "right" directions, respectively.

Two anchors were included among the systems to grade. One was the unprocessed signal reproduced by the acoustic pointer alone, and the other was a diffuse case where the five loudspeakers radiated the unprocessed signal convolved with uncorrelated 10 ms long sequences [22]. Each subject ran a training session before the actual test in order to familiarise with the different stimuli to be graded. The presentation order was randomised and the subjects did not know which system (or anchor) they were grading at any given time.

3.2. Results

The experimental setup described in Section 3.1 was replicated via simulations. Notice that, due to the symmetry of the setup in the central listening position, the simulated eardrum signals for the case of middle incidence direction are identical. This case is not included in the comparison. In fact, given identical signals, the model output would be equivalent to that of a natural source. The real subjects, on the other hand, could distinguish the two cases because their heads were not physically constrained [3].

The Pearson correlation between the subjective locatedness scores and the calculated localisation uncertainties H is -0.94, confirming that the model predictions are strongly correlated with the experimental data. The subjective scores and model predictions are compared side by side in Figure 4. In all cases except one (HOA in the central position) the model predictions are within the 95% confidence intervals of the experimental data. Similar results are obtained if the spherical head model is replaced with measured HRTFs [8]. Fine tuning of the p-norm distance is also not critical, as long as p is smaller than 1. In particular, values of p between 0.2 and 0.6 all give correlation coefficients close to -0.94. A correlation of -0.84 is obtained with the Manhattan distance. The Euclidean distance, on the other hand, yields a much weaker correlation of -0.56.

Note that the time of arrival difference between loudspeakers L_R and L_{SL} in the off-centre position is approximately 1.6 ms, which is above the 1 ms threshold where the law of the first front is in effect [3, 23]. The signal emitted by L_{SL} for the HOA system is not negligible. The predicted uncertainty is accurate in this case as well, although the proposed model (similarly to [6]) does not incorporate inhibitory mechanisms [3], which are thought to be at the base of the law of the first front [4]. This may not be true in general, and therefore the model predictions should be considered reliable only for arrival delays below 1 ms.

4. CONCLUSIONS

A computational model for the prediction of the localisation uncertainty was proposed. A novel way of using a joint representation of ILD and ITD cues was suggested. The mismatch between reference natural ILD-ITD pairs and those measured in experiments with multichannel audio systems was calculated using the 0.5-norm distance, and this choice was motivated on psychoacoustic grounds. On the basis of this distance, a score function was calculated, which was interpreted as the likelihood of an auditory event being perceived in a given direction. The localisation uncertainty was then calculated as the entropy of the normalised score function. The model was validated by means of comparison with experimental data. It was found that model predictions achieved a very high correlation with results of subjective listening tests.

5. REFERENCES

[1] Soren Bech and Nick Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*, John Wiley & Sons, 2006.

[2] L.A. Jeffress, "A place theory of sound localization," *J. of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35, 1948.

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

[4] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.*, vol. 80, pp. 1608, 1986.

[5] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling.," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 98–110, Jul. 1993.

[6] V. Pulkki and T. Hirvonen, "Localization of virtual sources in multichannel audio reproduction," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 13, no. 1, pp. 105–119, 2005.

[7] R.O. Duda and W.L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, pp. 3048, 1998.

[8] William G. Gardner and Keith D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995.

[9] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, 1993.

[10] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[11] L.R. Bernstein, S. van de Par, and C. Trahiotis, "The normalized interaural correlation: Accounting for nosπ thresholds obtained with gaussian and "low-noise" masking noise," *J. Acoust. Soc. Am.*, vol. 106, pp. 870, 1999.

[12] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, pp. 3075, 2004.

[13] Michel Marie Deza and Elena Deza, *Encyclopedia of distances*, Springer, 2009.

[14] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Trans. on Audio, Speech and Language Process.*, (submitted).

[15] ITU-R, *Recomm. BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems*, 2003.

[16] L.S.R. Simon and R. Mason, "Time and level localization curves for a regularly-spaced octagon loudspeaker array," presented at the AES 128th Conv., Preprint #8079, London, UK, May 2010.

[17] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," presented at AES 44th Conv., Preprint #C-4, Rotterdam, the Netherlands, March 1973.

[18] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Proc. AES 23rd Int. Conference, Copenhagen, Denmark*, May 2003.

[19] M. Poletti, "A unified theory of horizontal holographic sound systems," *J. Audio Eng. Soc.*, vol. 48, no. 12, pp. 1155–1182, Dec. 2000.

[20] J. Daniel, J.B. Rault, and J.D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," presented at the AES 105th Conv., Preprint #4795, San Francisco, California, USA, Sep. 1998.

[21] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Ph.D. thesis, University of Paris VI, France, 2000.

[22] G.S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, 1995.

[23] R.Y. Litovsky, H.S. Colburn, W.A. Yost, and S.J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, Oct. 1999.