This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Localization Uncertainty in Time-Amplitude Stereophonic Reproduction

Enzo De Sena, *Senior Member, IEEE*, Zoran Cvetković, *Senior Member, IEEE*, Hüseyin Hacıhabiboğlu, *Senior Member, IEEE*, Marc Moonen, *Fellow, IEEE*, Toon van Waterschoot, *Member, IEEE*

***Abstract*—This paper studies the effects of inter-channel time and level differences in stereophonic reproduction on perceived localization uncertainty, which is defined as how difficult it is for a listener to tell where a sound source is located. Towards this end, a computational model of localization uncertainty is proposed first. The model calculates inter-aural time and level difference cues, and compares them to those associated to free-field point-like sources. The comparison is carried out using a particular distance functional that replicates the increased uncertainty observed experimentally with inconsistent inter-aural time and level difference cues. The model is validated by formal listening tests, achieving a Pearson correlation of 0.99. The model is then used to predict localization uncertainty for stereophonic setups and a listener in central and off-central positions. Results show that amplitude methods achieve a slightly lower localization uncertainty for a listener positioned exactly in the center of the sweet spot. As soon as the listener moves away from that position, the situation reverses, with time-amplitude methods achieving a lower localization uncertainty.**

***Index Terms*—Stereophony, panning, recording and reproduction, localization uncertainty, auditory modelling.**

I. Introduction

DESPITE significant advancements in the field of multichannel audio [4], the most common reproduction system in use today remains the two-channel stereophonic system. In typical stereophonic panning, the two loudspeakers are positioned at $\pm 30^\circ$ with respect to the listener's look direction and reproduce delayed and attenuated versions of the same signals. The differences in time and level are typically frequency-independent, and are referred to as inter-channel time difference (ICTD) and inter-channel level difference (ICLD), respectively. For ICTDs smaller than 1 ms, the listener does not perceive the two loudspeaker signals as separate, but rather a single fused auditory event, often referred to as "phantom source". The perceived location of the phantom source depends on both the ICTD and ICLD. This psychoacoustic effect is called "summing localization", and is at the basis of stereophonic panning [5]. The phantom source can be moved using ICLDs alone (amplitude panning), ICTDs alone (time panning) or both ICLDs and ICTDs (time-amplitude panning).

Panning is inherently linked to recording. Consider a plane wave impinging on two microphones, each connected to a loudspeaker without mixing. The distance between the microphones dictates the ICTDs, while the ratio between the two directivity patterns (e.g. cardioid, figure-8) dictates the ICLDs. Recording with coincident microphones is equivalent to amplitude panning, while recording with non-coincident omnidirectional microphones is equivalent to time panning. Recording with microphones that are neither omnidirectional nor coincident is equivalent to time-amplitude panning. For point-like sources, distance attenuation means that small, distance-dependent level differences will also be observed at non-coincident microphones.

Amplitude panning is widely used in sound mixing, with most mixing desks and digital audio workstation (DAW) software implementing a version of the sine/tangent law [6], [7]. Amplitude recording methods are used in a wide variety of methods, e.g. the original Blumlein pair [8], Ambisonics [9], and the spatial decomposition method [10]. While it is possible to pan a stereophonic image using time panning for certain signals, it provides higher localization uncertainty for sustained higher frequency stimuli [11].

Time-amplitude recording methods [6], [8], on the other hand, are popular within the audio engineering community, mainly for their strong sense of spaciousness, which may be attributed to the higher decorrelation between the microphone signals [5], [6], [8]. Widely-spaced microphones (in the order of meters), while extensively used, are often criticized for their unstable imaging [12] and irregular distribution of reproduced auditory events [8]. Near-coincident stereophonic microphones such as the ORTF (*Office de Radiodiffusion Télévision Française*), NOS (*Nederlandse Omroep Stichting*) and DIN (*Deutsches Institut für Normung*) pairs have also been used widely in practice [13]. These arrays are preferred by practitioners for providing a stable and natural stereophonic image. ORTF was shown to have a localization curve most similar to a binaural recording [14]. A 3D extension to ORTF was recently proposed [15] and was shown to provide a good

Enzo De Sena is with the Institute of Sound Recording at the University of Surrey (UK) (e.desena@surrey.ac.uk). Zoran Cvetković is with the Department of Informatics at King's College London (UK). Hüseyin Hacıhabiboğlu is with the Graduate School of Informatics, Middle East Technical University (METU), Ankara, TR-06800, Turkey. Marc Moonen and Toon van Waterschoot are with the Department of Electrical Engineering at KU Leuven (Belgium). The work reported in this paper was partially funded by (i) EPSRC Grant EP/F001142/1, (ii) European Commission under Grant Agreement no. 316969 within the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)" (iii) European Research Council under the European Union's Horizon 2020 research and innovation program/ERC Consolidator Grant: SONORA (no. 773268), (iv) KU Leuven internal funds C2-16-00449 "Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking". This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Parts of this work were previously presented in [1], [2] and [3].

0000–0000/00$00.00 © 2019 IEEE

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

overall localization and auditory spaciousness in comparison with a coincident recording setup [16].

More recent work on time-amplitude recording techniques generally consider the problem from a constrained perspective, by using standard microphone directivities, by evaluating spatial acuity only at the sweet spot, or both. For example, a psychoacoustical evaluation of equal segment microphone array (ESMA) [17], revealed that the selection of the recording array dimensions has a distinct effect on the localization accuracy for an ESMA using cardioid microphones [18]. In another subjective study, first-order Ambisonics with *max-rE* encoding provided lower stereophonic image shifts in comparison with ESMA for a central listening position [19].

While the design of microphone arrays for recording spatial audio has traditionally been an *ad hoc* process driven by practical evidence that is not necessarily objectively validated, systematic approaches have also been proposed. A recently proposed design tool called microphone array recording and reproduction simulator (MARRS) allows the designer to design a stereophonic microphone pair using standard microphone directivity patterns (e.g. cardioid) and to observe the performance of the design both by visualization of the resulting localization curves and auralization of the simulation [20].

A similar systematic framework for the design of amplitude and time-amplitude circular multichannel recording and reproduction systems was proposed in [21], based on earlier work [22], [23]. An objective analysis based on active intensity fields showed that for stable rendition of plane waves it is beneficial to render each such wave by no more than two loudspeakers, thus re-framing the multichannel problem as a stereophonic one. Using available psychoacoustic curves, a family of optimal microphone directivity patterns was obtained, parametrized by the array radius. The obtained directivity patterns are too spatially selective to be implemented using first-order microphone patterns (e.g. hypercardiod), but can be implemented using higher-order microphones, e.g. differential microphones [24], [25]. Formal listening experiments were carried out for a microphone array with 15.5 cm radius [21]. Results showed a significantly improved localization accuracy with respect to Johnston's array, and comparable to that of vector-base amplitude panning (VBAP) and Ambisonics when the listener is in the center of the loudspeaker array. The experiments also assessed the localization uncertainty, defined as how difficult it is for the listener to tell where the sound source is located. Results showed an improvement in localization uncertainty with respect to VBAP and Ambisonics when the listener is in a position 30 cm off-center.

This paper explains why that is the case and shows that this is a more general characteristic of time-amplitude methods. Towards this end, this paper proposes a computational model of localization uncertainty (the model also allows prediction of the perceived direction, but this is left for future work).

Formal listening experiments require very careful design, and carrying them out is expensive and time-consuming [26]. Computational models provide a fast and repeatable alternative. Spatial hearing involves several stages of processing of sound waves impinging on the listener's head, which have been the subject of intensive study over the years. Well-established models now exist for the early stages of the mechanisms of spatial hearing, such as the effects of head diffraction, cochlear filtering and neural transduction. On the other hand, the higher levels of processing, where the spatial cues are combined, are not well understood yet. Various models have been proposed in the literature, but none has proved to be capable of predicting all characteristics of human hearing.

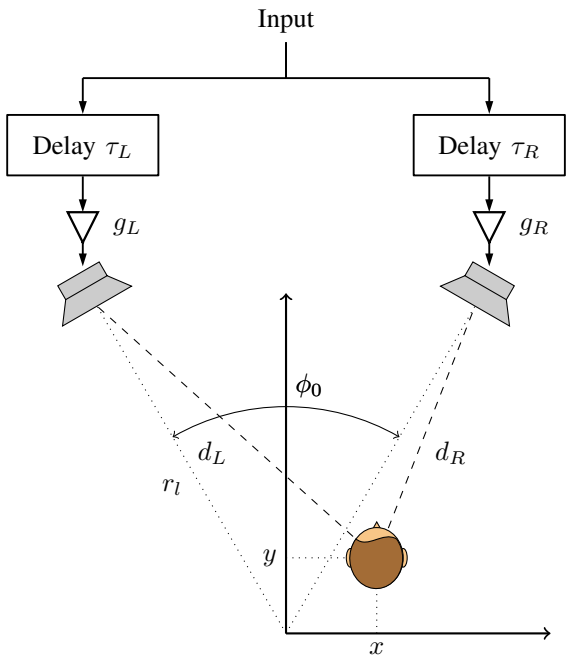


Fig. 1. Reference system for the stereophonic reproduction system.

This paper proposes a model that first calculates interaural time difference (ITD) and interaural level difference (ILD) cues that arise from stereophonic reproduction and then compares them to the ones associated to point-like sources. The comparison is carried out using a distance function which replicates the auditory event splitting observed with inconsistent ITD-ILD cues. It is shown that predictions of localization uncertainty based on this model are highly correlated with subjective scores. This model is then applied to generic ICTD-ICLD values in central and off-central positions to assess how time-amplitude stereophony affects localization uncertainty.

The paper is organized as follows. Section II provides the background on time-amplitude stereophony and on auditory system modelling. Section III presents the proposed model to predict localization uncertainty. Section IV discusses how time-amplitude panning affects localization uncertainty in stereophonic reproduction. Section V narrows the focus on a specific family of time-amplitude panning curves proposed in [21]. Section VI concludes the paper.

II. Background

A. Stereophonic Reproduction

Consider a stereophonic reproduction setup as shown in Fig. 1 with base angle ϕ_0 and loudspeaker distance r_l. The two loudspeakers are reproducing delayed and attenuated versions

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

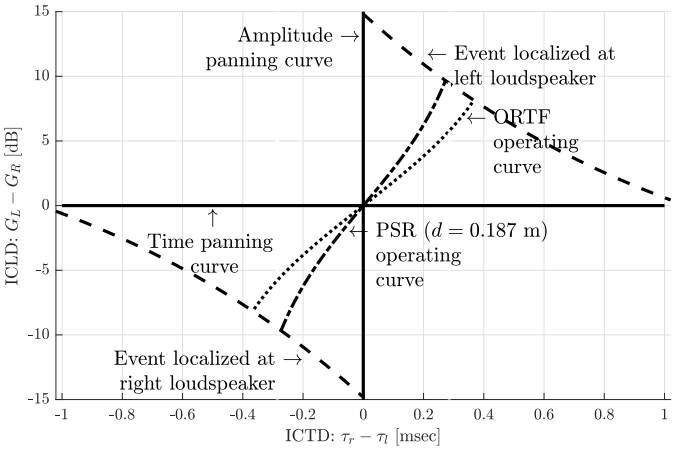


Fig. 2. Williams psychoacoustic curves, together with the panning curve associated to the PSR method with 18.7 cm inter-microphone distance and to the ORTF microphone pair. Amplitude methods are associated to points on the y-axis (ICTD $= 0$ ms).

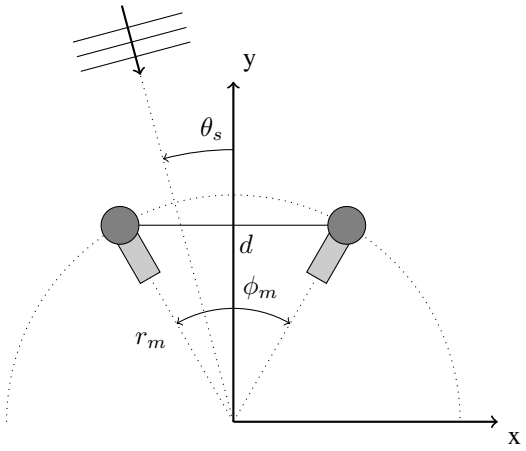


Fig. 3. Reference system for the stereophonic recording system with incoming plane wave.

of the same signal. The gains applied to the left and right loudspeaker are denoted as g_L and g_R, respectively, while the delays are τ_L and τ_R. The ICLD is defined as ICLD $= 20\log_{10}\frac{g_L}{g_R} = G_L - G_R$, where $G_L = 20\log_{10} g_L$ and $G_R = 20\log_{10} g_R$. The ICTD, on the other hand, is defined as $\tau_R - \tau_L$. Notice how these definitions are given such that whenever ICTD and ICLD have the same sign, their effect is consistent to one another. For instance, when both are positive, the left loudspeaker is louder and its signal also arrives earlier.

If the ICTD is below the *echo threshold*, the listener will perceive a single fused auditory event [5]. The echo threshold is strongly stimulus-dependent, and varies between 2 ms (for clicks) and 40 ms (for speech) [5]. For ICTDs between 1 ms and the echo threshold, the auditory event is localised at the loudspeaker whose signal arrives first [5], [27]. This effect is called "law of the first wavefront" [28]. For ICTDs smaller than 1 ms, a single fused "phantom" sound source is localised in a position that depends on both ICTD and ICLD. This psychoacoustic effect is called "summing localization" [5]. In the literature, summing localization and the law of the first wavefront are collectively referred to as "precedence effect" [5].

Fig. 2 shows Williams time-amplitude psychoacoustic curves, which represent all the ICLD-ICTD pairs that render the phantom source in the direction of the left and right loudspeaker [29]. Other time-amplitude psychoacoustic curves are also available in the literature [6], [30], [31].

The psychoacoustic curves show that if one wishes to render a phantom source in the direction of the left (respectively, right) loudspeaker it is possible to use an ICLD of about 15 dB (respectively, -15 dB) without any ICTD (amplitude panning). It is then possible to continuously pan between the two loudspeakers using ICLDs that vary between these two extremes (notice that the psychoacoustic curves do not give information on how to do this exactly, and only provide information about the extreme directions). Fig. 2 also shows that it is possible to render a sound source in the direction of the left loudspeaker by reducing the ICLD but increasing the ICTD (time-amplitude panning). If the ICLD is reduced to zero (time panning), it is still possible to displace the phantom source all the way to the left (respectively, right) loudspeaker with an ICTD of about 1 ms (respectively, -1 ms), which corresponds to the onset of the law of the first wavefront.

Note that most of the available stereophonic panning curves assume the standard stereophonic setup involving a phantom source panned between two loudspeakers positioned symmetrically to the left and right of the listener at a specified distance and can only provide suboptimal performance for panning lateral sources.

B. Stereophonic recording

Consider the reference system in Fig. 3. Here, the microphones are positioned on a circle with radius r_m facing outwards. The inter-microphone distance is denoted by d and is related to the array radius by

$$d = 2r_m \sin\left(\frac{\phi_m}{2}\right) , \tag{1}$$

where ϕ_m is the angle separating the two microphones, typically referred to as microphone base-angle. The value θ_s denotes the angle of an incoming plane wave. The arrangement of the two microphones on a circle (as opposed to positioning them on the x-axis) is preferred in this paper because it facilitates the extension to the multichannel case, and aids the comparison with PSR, as discussed later.

The first systematic approach to the problem of two-channel stereophonic recording is attributed to Blumlein [6], [8]. The Blumlein pair consists of two figure-8 microphones positioned orthogonally to each other ($\phi_m = 90°$) in the same position ($d = 0$ cm). Here, each microphone is connected to the corresponding loudspeaker without mixing, which means that the inter-microphone time and level differences are identical to ICTDs and ICLDs, respectively. Systems without mixing are also the focus of the rest of this paper. In the specific case of the Blumlein pair, the ICTDs are zero, since the microphone pair is coincident.

A variety of other stereophonic arrangements have been proposed in the past few decades (see e.g. [6], [8]). These

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Microphone arrangement	Inter-mic. distance d	Microphone base-angle ϕ_m	Directivity pattern	Coverage angle
PSR [21]	Adjustable	ϕ_0	Custom	ϕ_0
Blumlein [8]	0 cm	90°	Figure-8	68°
90-deg XY [8]	0 cm	90°	Cardioid	176°
ORTF [8]	17 cm	110°	Cardiod	94°
DIN [8]	20 cm	90°	Cardioid	100°
NOS [8]	30 cm	90°	Cardioid	80°

TABLE I
CHARACTERISTICS OF POPULAR MICROPHONE ARRANGEMENTS. HERE, THE COVERAGE ANGLE DENOTES THE RANGE OF ANGLES THAT RESULTS IN ICTD-ICLD PAIRS WITHIN THE WILLIAMS CURVES.

include the ORTF, DIN, NOS and cardioid XY pair, the characteristics of which are summarised in Table I. The operating curve associated to the ORTF is shown in Fig. 2.

C. Perceptual Soundfield Reconstruction

The methods discussed in the previous subsection were mostly designed on a trial-and-error basis. A systematic framework for the design of circular microphone arrays was proposed in [21], based on earlier work by Johnston and Lam [22] and termed perceptual sound-field reconstruction (PSR). This section summarizes the design procedure, but for the specific case of a stereophonic setup.

In PSR, each microphone is connected to a corresponding loudspeaker. The microphone base-angle, ϕ_m, is identical to the loudspeaker base-angle, ϕ_0, which allows a straightforward extension to full 360° multichannel rendering. The inter-microphone delay (and thus ICTD) is:

$$\text{ICTD}(\theta_s) = \frac{d}{c}\sin\theta_s = \frac{1}{c}2r_m \sin\left(\frac{\phi_0}{2}\right)\sin\theta_s \ , \qquad (2)$$

where c is the speed of sound ($c = 343$ m/s in dry air at 20°). Notice that in the PSR literature, the ICTD is expressed as a function of the array radius, r_m. The remainder of this paper, on the other hand, uses the inter-microphone distance, d, which facilitates the comparison with other popular stereophonic microphone arrangements.

Consider the ICTD obtained for a plane wave with $\theta_s = \frac{\phi_0}{2}$, i.e. in the direction of the left microphone:

$$\text{ICTD}\left(\phi_0/2\right) = \frac{d}{c}\sin\left(\frac{\phi_0}{2}\right) \ . \qquad (3)$$

Given the value of $\text{ICTD}\left(\phi_0/2\right)$, psychoacoustic time-amplitude curves can be used to find the minimum ICLD necessary to render a phantom source in the direction of the left loudspeaker. Rather than using Franssen curves as in [21], which are considered to be not sufficiently precise for quantitative design [5], this paper uses Williams curves. Setting the ICLD to be equal to the Williams value yields the constraint $\text{ICLD}\left(\phi_0/2\right) = \text{ICLD}_{W,L}(\text{ICTD}\left(\phi_0/2\right))$, where $\text{ICLD}_{W,L}(\text{ICTD})$ is the Williams curve associated to the left loudspeaker (the top curve in Fig. 2). The value $\text{ICLD}_{W,L}(\text{ICTD}\left(\phi_0/2\right))$ will denoted by ICLD_W in the following. For $\theta_s = -\frac{\phi_0}{2}$ (i.e. the direction of the right loudspeaker) one has $\text{ICLD}\left(-\phi_0/2\right) = \text{ICLD}_{W,R}(\text{ICTD}\left(-\phi_0/2\right))$, which, due to the symmetry of the problem, is equivalent to $\text{ICLD}\left(-\phi_0/2\right) = -\text{ICLD}_W$. A third trivial point can be added, i.e. $\text{ICLD}(0) = 0$ dB.

It remains to choose how to connect these three points. Two choices were explored in [21]: a simple straight line [32], or a modified version of the tangent panning law [21]. While both approaches were shown to lead to a good localization accuracy (i.e. the phantom source was shown to be perceived close to the intended direction θ_s), the latter approach allows to link the design with other methods (e.g. VBAP [7]) and consists of the following parametric function:

$$\text{ICLD}(\theta_s) = 20\log_{10}\frac{\sin\left(\frac{\phi_0}{2}+\beta+\theta_s\right)}{\sin\left(\frac{\phi_0}{2}+\beta-\theta_s\right)} \ , \qquad (4)$$

where β is a free parameter that is used to satisfy the constraint $\text{ICLD}(\phi_0/2) = \text{ICLD}_W$ (the other two points are then also satisfied due to symmetry), which results in

$$\beta = \arctan\left(\frac{10^{\frac{\text{ICLD}_W}{20}}\sin(\phi_0)}{1-10^{\frac{\text{ICLD}_W}{20}}\cos(\phi_0)}\right) \ . \qquad (5)$$

To summarize, in order to obtain the PSR panning curve one should (a) set the free parameter d (or, equivalently, r_m), (b) obtain $\text{ICTD}\left(\phi_0/2\right)$ from (3), (c) obtain ICLD_W from Williams psychoacoustic curve and β from (5) and (d) obtain $\text{ICTD}(\theta_s)$ and $\text{ICLD}(\theta_s)$ from (2) and (4), respectively. The result of this procedure is a family of panning curves parametrized by the value of the inter-microphone distance d. In the extreme case $d = 0$, one obtains an amplitude-only method ($\text{ICTD}(\theta_s) = 0 \ \forall \ \theta_s$). As the value of d increases, the ICTDs increase while the ICLDs dictated by (4) decrease, thus achieving stereophonic rendering with a different time/amplitude trading ratio. Fig. 2 shows the panning curve obtained for $d = 18.7$ cm.

The so-obtained $\text{ICLD}(\theta_s)$ and $\text{ICTD}(\theta_s)$ can be viewed either in the context of stereophonic panning, whereby they can be used to directly control appropriate loudspeaker gains and delays, or in the context of stereophonic recording. In the latter case, the microphone directivity patterns can be designed so as to emulate $\text{ICLD}(\theta_s)$. Thus, one wishes to set $\text{ICLD}(\theta_s) = 20\log_{10}\frac{\Gamma_L(\theta_s)}{\Gamma_R(\theta_s)}$, where $\Gamma_L(\theta_s)$ and $\Gamma_R(\theta_s)$ denote the directivity patterns of the left and right microphones. In general, first-order microphones (i.e. microphones with $\Gamma(\theta_s)$ of the type $\Gamma(\theta_s) = a_0 + a_1\cos(\theta_s)$) are not sufficiently directive to achieve the necessary ICLDs. Second-order microphones are already sufficient for this purpose [21]. The remainder of this paper will refer to PSR *panning/recording* to emphasize the fact that application is possible in both contexts.

Notice that PSR aims at panoramic reproduction where a listener is allowed not only to move their head (e.g. to reduce localization ambiguities) but also to freely rotate their orientation with respect to an arbitrary, reference reproduction axis. In order to provide an orientation-independent homogeneous localization for a symmetric loudspeaker setup, stereophonic pairwise panning laws can be used as a starting point. Despite the fact that these laws provide suboptimal localization performance for lateral sources, sources that are positioned within the range of near-peripheral azimuths (i.e. $\pm\phi_0/2$ with respect

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

to the listener's frontal direction) can be rendered accurately. From a purely physical point of view, PSR was also shown, via an energetic analysis of the reproduced sound field, to provide a good reproduction of directional properties of the sound field in a wide listening area [21].

D. Auditory system modelling

The auditory system estimates the directions of sound sources based on a combination of monaural and binaural cues [5]. Localization in the horizontal plane is mostly reliant on binaural cues, particularly on differences in the time of arrival and on difference in level of a sound wave at the two ears. ITDs are caused by the different time of arrival of sound waves radiated by sources outside the median plane. At low frequencies the auditory system analyses the interaural time difference between the signals' fine structure [5]. At higher frequencies this mechanism becomes ambiguous, and the time differences between the signals' envelopes are used instead [5]. The maximum naturally occurring ITD is approximately 0.65 ms [5], [6]. ILDs are caused by the acoustical shadowing of the head and are strongly frequency-dependent. At low frequencies the head is approximately transparent to the sound wave and the level differences are small. As the wavelength approaches the size of the human head, the level differences become sensible. The highest natural ILD is in the region of 20 dB [5].

A common confusion in this context is to assume that ICTDs are identical to ITDs, and ICLDs are identical to ILDs, which is incorrect. The wavefronts of each loudspeaker reach both ears and form interference patterns at the position of the ears, which, in turn, lead to a complex relationship between ICTD-ICLD and ITD-ILD. Also notice that, as opposed to ICTDs and ICLDs, ITDs and ILDs are frequency-dependent.

The mechanisms by which the auditory system interprets the ITD and ILD cues are complex and not yet fully understood [5]. Experimental evidence suggests that humans use two main mechanisms for source localization, and that these mechanisms are to a certain degree independent from one another [5, 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*. The latter mechanism seems to be dominant for signals with significant frequency content above 1.6 kHz [5, p.173].

A first, most notable attempt to model binaural processing was made by Jeffress in 1948 [33], who hypothesised that sound localization is governed by a mechanism of running cross-correlation between the two channels. While today this is still considered to be an adequate mean of measuring ITDs, it does not account for the presence of ILDs [5]. Lindemann [34] proposed a model that incorporates this information in the cross-correlation mechanism by way of inhibitory elements that are physiologically plausible. Gaik [35] extended this model further, based on the observation that ITDs and ILDs due to point-like sources in free field come in specific pairs. For instance the ITD and ILD values for a source in the median plane are both small. On the other hand, for a source to the right/left, both ITD and ILD are high. In fact, in these cases the acoustic wave arriving at the far ear is both attenuated (because of head shadowing) and delayed (because of propagation time).

Gaik observed that when inconsistent ITD-ILD pairs (e.g. a left-leading ILD and a right-leading ITD) are presented over headphones, the auditory event width increases, and sometimes two separate events appear [5], [35]. In other words, inconsistent ITD-ILD pairs cause increased localization uncertainty. These unnatural conditions can arise also with multiple sources radiating coherent signals, as in stereophonic reproduction. Indeed, although each loudspeaker acts as a free-field source, the signals due to the different loudspeakers add up at the ears, creating interference phenomena that may result in inconsistent ITD-ILD cues. Quantifying the deviation between the reproduced ITD-ILD pairs and the ones associated to natural sources is therefore useful to study the localization uncertainty due to different multichannel methods. A study presented by Pulkki and Hirvonen in [36] goes in this direction. For a given multichannel method they find the angle of the closest free-field source in terms of ILD and ITD, separately. This model gives useful predictions when the angles corresponding to the ILD and to the ITD coincide. However, in most cases the ITD and ILD cues provide contradicting information, and therefore the model output is hard to interpret [36].

III. Localization Uncertainty Model

The first step of the model is to calculate ITD-ILD pairs of single point-like free-field sound sources in a number of directions on the horizontal plane. The so-obtained pairs are referred to as *free-field ITD-ILD pairs*. Similarly to [36] and [37], it is hypothesized here that the auditory system uses the free-field ITD-ILD pairs as a dictionary to interpret all other acoustical conditions. The ITD-ILD pairs for the acoustical scene to be estimated are calculated and compared to the free-field ITD-ILD pairs using a given distance functional. Finally, the information is combined across critical bands to obtain an overall estimate of the localization uncertainty.

A. Calculation of ILDs and ITDs

The ITD and ILD values are calculated as follows. Acoustic sources are modelled as point sources in the free field. The head related transfer functions (HRTFs) are taken as the Kemar mannequin measurements from the CIPIC database (subject 25) [38]. The sampling frequency is 44.1 kHz. The response of the cochlea is modelled using a gammatone filter-bank [39] with 24 center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale between 60 kHz and 15 kHz [40]. As a rough model of the neuron firing probability, the bandpass signals are half-wave rectified below 1.5 kHz, while above 1.5 kHz the envelope of each bandpass signal is taken using the discrete Hilbert transform [35]. The resulting signals are fed to 24 binaural processors that calculate the ITD and ILD values independently. The ITD is calculated as the location of the maximum of the cross-correlation function evaluated over time lags between $[-0.7, 0.7]$ ms [35], [37]. The ILD is calculated as the energy ratio of the left and right

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

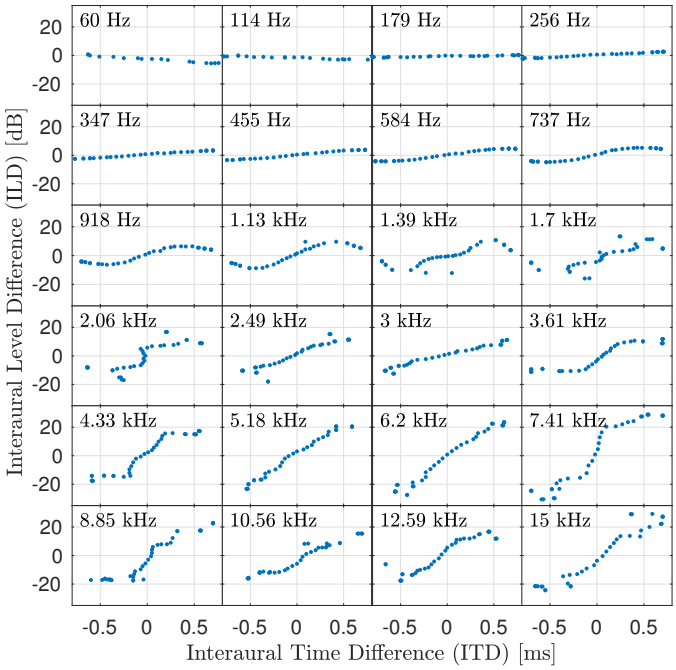


Fig. 4. The figure shows the ITD-ILD pairs associated to point-like free-field sources in each critical band.

channel [37]. Altogether, the model produces a set of 24 ITD-ILD pairs (48 values in total).

Fig. 4 show the ITD-ILD pairs associated to free-field sources in the frontal horizontal plane. The free-field sources are reproducing 50 ms long white noise sample, multiplied by a Tukey window with taper parameter 5%. Each simulation is repeated ten times to average out the effect of different noise realizations. Each point on the plot corresponds to a free field source positioned at angles between $\theta = -90°$ and $\theta = 90°$ with an angular resolution of $5°$. In the remainder of this paper, these values will be denoted as $\text{FITD}_i(\theta)$ and $\text{FILD}_i(\theta)$, respectively, where i is the critical band index and θ is the free-field source angle with respect to the listener's look direction.

It can be observed in Fig. 4 that the interaural cues are highly correlated, i.e. larger ITD values are typically associated to larger ILD values, which is due to the concurrent effect of sound propagation and diffraction around the head [35]. Also, the maximum ILD values increase with frequency, which is due to the increasing head shadowing associated to decreasing wave lengths [35]. Some small asymmetries are observed, which are possibly due to noise or asymmetries in the measurement setup of the HRTF dataset.

B. Distance between ITD-ILD pairs

Let ILD_i and ITD_i denote the ITD and ILD values in the i-th critical band as observed by a listener under stereophonic reproduction (or other acoustical conditions).

In order to combine the information of ITD and ILD cues across critical bands, it is useful to normalize all quantities to the maximum values of the free-field cues:

$$\overline{\text{ITD}}_i = \frac{\text{ITD}_i}{\max_\theta |\text{FITD}_i(\theta)|} \ , \tag{6}$$

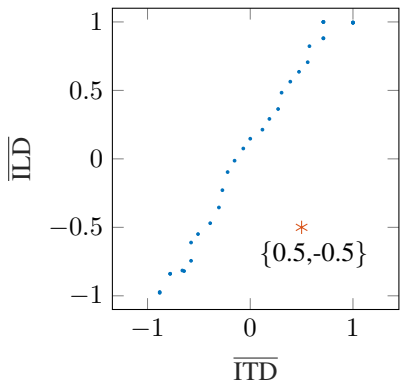


Fig. 5. The figure shows the normalized ITD and ILD pairs associated to free-field sources for the 18th critical band (black dots), and an example of a $\{\overline{\text{ITD}}, \overline{\text{ILD}}\}$ point in $\{0.5, -0.5\}$.

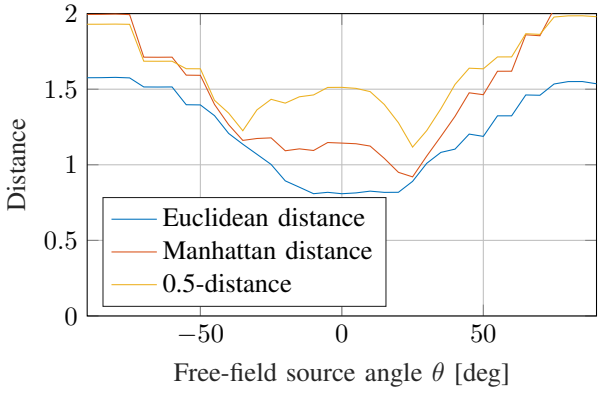


Fig. 6. Distance between $\{\overline{\text{ITD}}_{18}, \overline{\text{ILD}}_{18}\} = \{0.5, -0.5\}$ and the free-field point $\{\overline{\text{FITD}}_{18}(\theta), \overline{\text{FILD}}_{18}(\theta)\}$ as a function of direction θ and for different distance functionals.

$$\overline{\text{ILD}}_i = \frac{\text{ILD}_i}{\max_\theta |\text{FILD}_i(\theta)|} \ . \tag{7}$$

The free-field pairs, $\text{FITD}_i(\theta)$ and $\text{FILD}_i(\theta)$, are normalized in the same way and are denoted as $\overline{\text{FITD}}_i(\theta)$ and $\overline{\text{FILD}}_i(\theta)$, respectively.

Notice that while $\overline{\text{FITD}}_i(\theta) \in [-1, 1]$ and $\overline{\text{FILD}}_i(\theta) \in [-1, 1]$, $\overline{\text{ITD}}_i$ and $\overline{\text{ILD}}_i$ can in theory be outside that range. Indeed, there is no guarantee that the acoustical condition being analyzed has ITD-ILD values outside the range of values occurring for free-field sources.

The aim is now to select a distance functional between $\{\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\}$ and $\{\overline{\text{FITD}}_i(\theta), \overline{\text{FILD}}_i(\theta)\}$ according to some meaningful psychoacoustic criterion. Consider the distance defined by the classical p-norm:

$$\left(\left|\overline{\text{ITD}}_i - \overline{\text{FITD}}_i(\theta)\right|^p + \left|\overline{\text{ILD}}_i - \overline{\text{FILD}}_i(\theta)\right|^p \right)^{\frac{1}{p}} \ . \tag{8}$$

Fig. 5 shows an example of an observed $\{\overline{\text{ITD}}_{15}, \overline{\text{ILD}}_{15}\}$ point positioned at $\{0.5, -0.5\}$ in the 15th critical band (the one centered at 6.07 kHz). The distance between each free-field source $\{\overline{\text{FITD}}_{15}(\theta), \overline{\text{FILD}}_{15}(\theta)\}$ and the $\{\overline{\text{ITD}}_{15}, \overline{\text{ILD}}_{15}\} = \{0.5, -0.5\}$ point is plotted in Fig. 6 as a function of the free-field source angle, θ, for different distance functionals.

The experimental evidence shows that subjects presented

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

with contradicting ITD-ILD pairs are likely to report split auditory events [35]. The Euclidean distance ($p = 2$) does not emulate this behaviour, as it leads to a single minimum in $\theta = 0$, as shown in Fig. 6. The Manhattan distance ($p = 1$) is nearly constant in the $[-30°, 30°]$ angular sector. The 0.5-distance, on the other hand, causes two sharp minima, one of which is centred in the direction corresponding to the ITD cue, which is compatible with the psychoacoustic evidence [5, p.170]. Other values of p close to 0.5 would also retain this behaviour. In the next section it is shown that the model has very good predictive power even without careful tuning of p.

The distance defined in (8) does not satisfy the triangle inequality for $p < 1$. However, the same distance raised to power p, does satisfy all the properties of a distance [41]. Hence, the p-norm distance used here is:

$$\xi_i\left(\theta|\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\right) = \left|\overline{\text{ITD}}_i - \overline{\text{FITD}}_i(\theta)\right|^p + \left|\overline{\text{ILD}}_i - \overline{\text{FILD}}_i(\theta)\right|^p . \tag{9}$$

The objective is to obtain a function of θ that quantifies the likelihood that a sound is perceived in that direction. Intuitively, this function should be inversely proportional to $\xi_i(\theta|\overline{\text{ITD}}_i, \overline{\text{ILD}}_i)$, such that whenever the distance is small, the likelihood is high (and viceversa). Let this function be

$$f_i\left(\theta|\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\right) = Ke^{-\xi_i\left(\theta|\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\right)} , \tag{10}$$

where K is a positive constant. Although other choices are available [1], the advantage of (10) is that it gives the model an explicit statistical interpretation in the maximum likelihood (ML) framework, as will be discussed in the next subsection.

The next step is to integrate the information from the different critical bands. The mechanisms governing this stage of perception are generally regarded as complex and not well understood [5], [36]. Here, the information across critical bands is combined as a loudness-weighted average:

$$f\left(\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right) = \frac{K}{N}\sum_{i=1}^{N} w_i e^{-\xi_i\left(\theta|\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\right)} , \tag{11}$$

where w_i are the loudness weights, $\overline{\text{ILD}}$ denotes the vector $\overline{\text{ILD}} = \left[\overline{\text{ILD}}_1, ..., \overline{\text{ILD}}_N\right]$ and $\overline{\text{ITD}} = \left[\overline{\text{ITD}}_1, ..., \overline{\text{ITD}}_N\right]$, and N is the number of critical bands ($N = 24$). The loudness weights are set according to the procedure proposed in [42], which results in critical bands with a higher signal level weighting more than critical bands with little active content [42]. The procedure involves (a) calculating the SPL levels in each critical band, (b) converting them to phon levels through the BS ISO 226:2003 equal-loudness contours, and (c) converting them to w_i weights using a function where a 10 phon reduction leads to a halving of the weight, in line with Stevens's model [43].

C. Statistical interpretation

Suppose that $\{\overline{\text{ITD}}_i, \overline{\text{ILD}}_i\}$ are noisy observations of a point-like free-field source at angle θ_0 associated to a set of true $\{\overline{\text{FILD}}_i(\theta_0), \overline{\text{FITD}}_i(\theta_0)\}$ values:

$$\begin{aligned} \overline{\text{ITD}}_i &= \overline{\text{FITD}}_i(\theta_0) + u_i , \\ \overline{\text{ILD}}_i &= \overline{\text{FILD}}_i(\theta_0) + v_i . \end{aligned} \tag{12}$$

where u_i and v_i are the noise components, which may arise for instance as a consequence of room reverberation or reproduction with multiple loudspeakers. Let the joint distribution of the noise components be the following mixture of zero-mean bivariate theta-generalized normal distributions [44]:

$$f\left(u_1, ..., u_N; v_1, ..., v_N\right) = \frac{1}{N}\sum_{i=1}^{N} Ce^{-\frac{|u_i|^p + |v_i|^p}{2\sigma^2}} , \tag{13}$$

where σ represents a standard deviation and C is a normalization constant.

Notice that if $\frac{1}{2\sigma^2} = 1$, equation (13) becomes identical to (11). Suppose now that the auditory system estimates the true value θ_0 using an ML approach. Within this context, $f\left(\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)$, seen here as a function of θ, takes the meaning of a likelihood function.

Although the objective of the proposed model is not to estimate θ itself (which is left for future work), the shape of the likelihood function gives information about how difficult it is to estimate it, as will be explained in the next subsection.

D. Calculation of the localization uncertainty

The likelihood function quantifies the probability that a subject would perceive a source in a given direction θ. From this perspective, a uniform (constant) likelihood function would result in a maximally uncertain (diffuse) event. At the other extreme, an impulsive likelihood function would result in a minimally uncertain event. Various measures can be used in this context. This paper uses a modified version of the circular variance [45], [46]. This measure is drawn from the field of directional statistics [46], and has values between 0, which is associated to an impulsive function, and 1, which is associated to a constant function. Calculating the modified circular variance involves normalising $f\left(\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)$ so that it sums to one (e.g. by adjusting the K constant), and calculating the (modified) first cosine and sine moments:

$$\alpha = \int_0^{2\pi} f\left(2\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)\cos(\theta)d\theta, \tag{14}$$

$$\beta = \int_0^{2\pi} f\left(2\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)\sin(\theta)d\theta , \tag{15}$$

where the modification consists of using $f\left(2\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)$ instead of $f\left(\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)$. This modification is motivated by the fact that the likelihood function only takes values in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, which would not lead to a circular variance of 1 in case of a constant likelihood function. Then, the circular variance is given by [46]

$$\text{H}\left(\overline{\text{ITD}}, \overline{\text{ILD}}\right) = 1 - \sqrt{\alpha^2 + \beta^2} . \tag{16}$$

The final step is to normalize the value of H. Indeed, since $f\left(\theta|\overline{\text{ITD}}, \overline{\text{ILD}}\right)$ is never impulsive (even for free-field sources), the value of H is always larger than zero. For instance, for $p = 0.7$, the minimum H is 0.49. The following normalization yields values close to zero for estimates associated to free-field sources:

$$\overline{\text{H}}\left(\overline{\text{ITD}}, \overline{\text{ILD}}\right) = \frac{\text{H}\left(\overline{\text{ITD}}, \overline{\text{ILD}}\right) - \text{H}_{\text{min}}}{1 - \text{H}_{\text{min}}} \tag{17}$$

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

where $\overline{\mathrm{H}}\left(\overline{\mathrm{ITD}}, \overline{\mathrm{ILD}}\right)$ denotes the final localization uncertainty estimates of the proposed model and $\mathrm{H}_{\mathrm{min}} = \min_\theta \mathrm{H}\left(\overline{\mathrm{FILD}}(\theta), \overline{\mathrm{FITD}}(\theta)\right)$.

It should be noted that, in practice, all calculations above are made using a discretization of the angles θ. The simulations in the remainder of this paper use a resolution of 5°.

E. Model validation

A formal listening test with 19 subjects was carried out in [21] using a modified MUSHRA test [47]. 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 test was carried out in an audio booth using four synthesized 5-channel surround sound methods: (a) pair-wise tangent panning law [48] (equivalent to horizontal VBAP); (b) near-field corrected second-order Ambisonics with mode-matching decoding at low frequency and maximum-energy decoding at high frequency [9], [49], [50]; (c) second-order Ambisonics with in-phase decoding [51]; and (d) the quasi-coincident microphone array proposed in [21]. The test was run for three sound source directions and two seating positions, and included both a reference (a loudspeaker in the intended direction) and an anchor (an approximately diffuse soundfield). The rendered virtual source directions were in front of the listener and $\pm 18^\circ$ to the right and left of the front direction. Three excerpts (female speech, african bongo and cello) were used as representatives of common program material, and the resulting subjective scores were averaged across the three excerpts. Details of the experiment are available in [21].

In order to validate the proposed model, the experiment was replicated here through simulations. The stimuli used in the simulations were (a) long white noise burst (500 ms), (b) short white noise burst (50 ms), (c) short pink noise burst (50 ms), and (d) impulsive sound. All stimuli (except for the impulsive sound) were multiplied by a Tukey window with a 5% taper parameter.

The loudspeaker signals for each of the 5-channel surround sound methods mentioned above were then calculated. Notice that the experiment and the simulations employed multichannel sound reproduction where more than two loudspeakers can be substantially active in some of the methods used (i.e. Ambisonics). While the model will be used in a stereophonic context in this paper, it is independent from the number of loudspeakers that the input ear signals originate from.

The acoustic path from each loudspeaker to the head was simulated in free-field conditions using appropriate time delays and distance attenuation (inverse square law) that depend on the position of the head. The ear pressure signals were obtained using the Kemar mannequin HRTFs. The so-obtained ear pressure signals associated with multiple loudspeakers were added together. Finally, the resulting ear pressure signals were fed to the proposed model, and localization uncertainty estimates were obtained using (17).

Fig. 7 shows the absolute Pearson correlation between the subjective scores and the localization uncertainty estimates, as a function of the parameter p for different stimuli. All the types of stimuli result in an absolute Pearson correlation stronger than -0.98 for $p \in [0.7, 0.9]$, confirming that the model predictions are strongly correlated with the experimental data. The reason why the correlation is negative is that subjects' answers to the question "How certain are you of the direction of the source?" are merely inverted with respect to localization uncertainty.

All types of random noise perform particularly well. This indicates that the choice of stimulus does not appear to be critical when aiming to predict the localization uncertainty of an experiment involving a varied programme material like the one in [21] (which included speech, bongos and cellos). Furthermore, since the free-field pairs $\{\overline{\mathrm{FITD}}_i(\theta), \overline{\mathrm{FILD}}_i(\theta)\}$ were obtained using short white noise bursts, the model appears to be robust to a mismatch between the stimulus used for the free-field pairs and the one used in the acoustic scene under analysis (in a machine learning context, these would be akin to training and testing, respectively).

Fine tuning of the p-norm distance is also not critical as long as contradicting ITD-ILD cues lead to a an approximately bimodal likelihood function. For the short white noise burst, values of p between 0.5 and 1.0 all give correlation coefficients of approximately -0.99. The Euclidean distance ($p = 2$), on the other hand, yields a much weaker correlation of -0.84. At the other extreme, a small p also leads to weak correlations, indicating that norms closer to the l_0-norm do not provide an effective measure in this context.

The simulations in the remainder of this paper use the short white noise burst, and a distance function with $p = 0.7$ (an approximate midpoint of the $p \in [0.5, 1.0]$ interval with -0.99 correlation). Fig. 8 shows the scatter plot comparing the model estimates with the subjective scores.

Note that the selection of an earlier set of results from an experiment using a five channel setup for the validation of the proposed model is deliberate. While the derivation of the model is based on two-channel stereophony as the simplest possible spatial audio reproduction system, its predictive power extends beyond that as is shown.

Note also that the experiment did not use any head or position tracking and the listeners were instructed to keep as still as possible. Such a listening scenario is not entirely ecologically valid. However, since the proposed model is neither dynamic nor does it incorporate any mechanism to account for listener motion, the results from the experiment are indicative of the predictive power of the model.

IV. Localization Uncertainty in Stereophonic Reproduction

A. Localization uncertainty in the center of the sweet-spot

Fig. 9 shows the localization uncertainty produced by the proposed model for a listener in the center of the sweet-spot and for a stereophonic reproduction system with base angle $\phi_0 = 60^\circ$ and $r_l = 2$ m (see Fig. 1). Two areas in the second and fourth quadrants have a significant localization uncertainty. These areas correspond to cases where ICLD and ICTD provide inconsistent information: one loudspeaker is leading in terms of ICLD (i.e. it is louder) while the other is leading in terms of ICTD (i.e. it arrives earlier). This indicates that

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

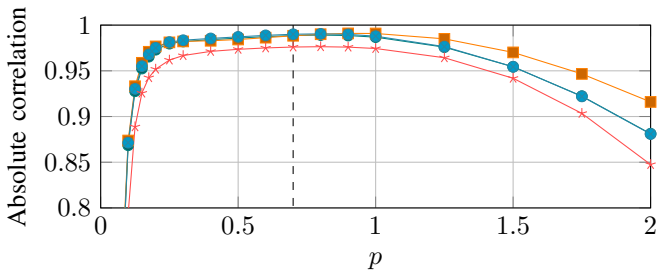


Fig. 7. Absolute Pearson correlation between the model predictions and the testing set as a function of the distance functional parameter p for different stimuli: short white noise burst (—●—), long white noise burst (—●—), impulse (—✶—) and pink noise (—■—). The short white noise burst curve (—●—) is not visible because it coincides almost exactly with the long white noise burst curve (—●—). The dashed vertical line denotes the value chosen in the remainder of the simulations in this paper ($p = 0.7$).

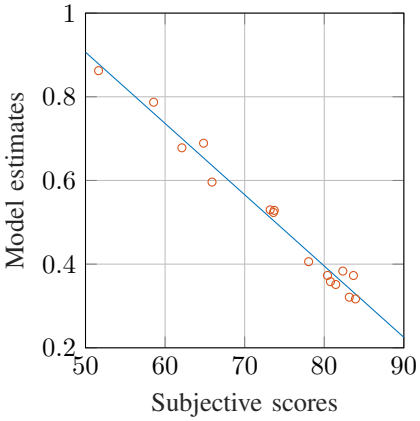


Fig. 8. Scatter plot comparing the subjective scores of the experiment presented in [21] and the estimates of the proposed model for $p = 0.7$ and for short white noise burst. The blue line indicates the best-fitting line. The Pearson correlation is -0.99.

inconsistent ICTD-ICLD pairs somehow translate to unnatural ITD-ILD cues, which is in agreement with the experimental findings of Leakey in [52]. It may also be observed in Fig. 9 that large ICLD values (outside $\approx \pm 13$ dB) result in a low localization uncertainty for all ICTDs. Here, one loudspeaker signal is masking the other.

Finally, it can be observed that for ICTDs around ± 0.3 ms the localization uncertainty increases even in the first and third quadrants, where ICTDs-ICLDs pairs are consistent. At ICTDs of approximately ± 0.3 ms, the two loudspeaker signals arrive at the same time at one of the two ears (more specifically, at the left ear for ICTD $\approx +0.3$ ms and at the right ear for ICTD ≈ -0.3 ms). Appendix A proves that these ICTDs can be approximated as

$$\tau_o \approx \pm \frac{r_h}{c} \left[\cos\left(\theta_e - \frac{\phi_0}{2}\right) + \frac{\phi_0}{2} + \theta_e - \frac{\pi}{2} \right], \qquad (18)$$

where r_h denotes the head radius and θ_e denotes the angle between the forward-looking direction and the ear. For $r_h = 9$ cm, $\theta_e = 100°$ and $\phi_0 = 60°$, then $\tau_o = \pm 0.27$ ms. If signals of both loudspeakers arrive at the same time at one of the ears, that ear effectively receives one instance of the rendered acoustic event, whereas the other ear receives two

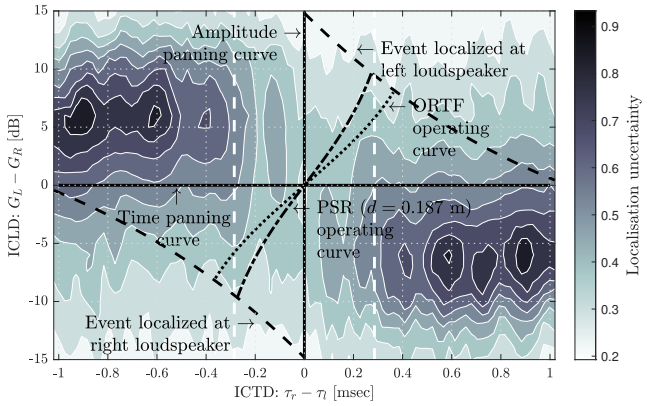


Fig. 9. Contour plot of localization uncertainty as a function of ICLD and ICTD in the center of the sweet-spot. The dashed white lines denote delays that result in the loudspeaker signals arriving at the same time at one of the two ears. Overlaid are the Williams' curves and the time-amplitude panning curve associated to the PSR method with $d = 18.7$ cm.

instances. A number of psychoacoustic studies investigated effects of presenting three coherent stimuli, two to one ear and the third to the other ear. It was found that the perceived event had "complex spatial structure" including cases where, depending on relative delays between the three stimuli, two distinct acoustic sources could be perceived [5]. The results of the proposed model are in agreement with the findings of these experiments. In fact, simulations not shown here for space reasons, confirm that as the head radius r_h changes, the areas with higher ICTDs in the first and third quadrants move in accordance to (18). If one wishes to avoid these areas, the ICTD values should be restricted to the open interval ICTD $\in]-\tau_o, \tau_o[$. In Fig. 9, notice how the PSR operating curve associated to $d = 18.7$ cm avoids the areas with higher localization uncertainty around ICTD $\approx \pm 0.3$ ms. This is so by construction, as will be discussed later in Section V-A.

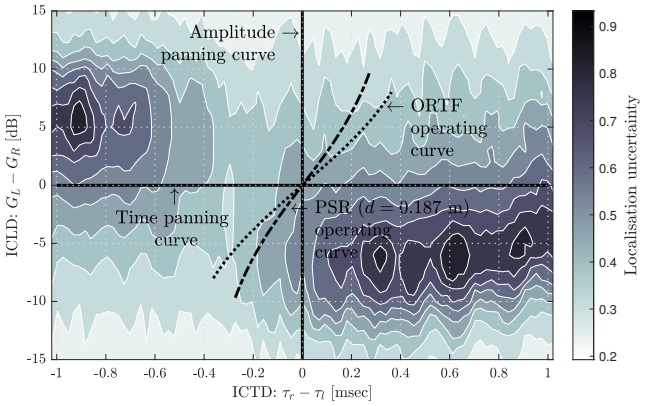
B. Localization uncertainty in off-center positions

Figure 10 shows the localization uncertainty for a listener in a position 10 cm to the left, and in a position 10 cm and 20 cm to the right of the sweet-spot. The listener is still looking ahead, in a direction parallel to the y-axis. It may be observed that these plots are almost identical to the on-center plot of Fig. 9 but shifted horizontally. This implies that the dominant effect is the change in the ICTDs observed by the listener as a consequence of having moved closer to the right loudspeaker. The change in observed ICLDs, on the other hand, appears to have a minor effect. Likewise, the relative change of direction of the loudspeakers also appears to have a minor effect (in the 0.2 m position, the two loudspeakers appear at 25° and $-35°$ with respect to the listener, compared to $\pm 30°$ for the central position).

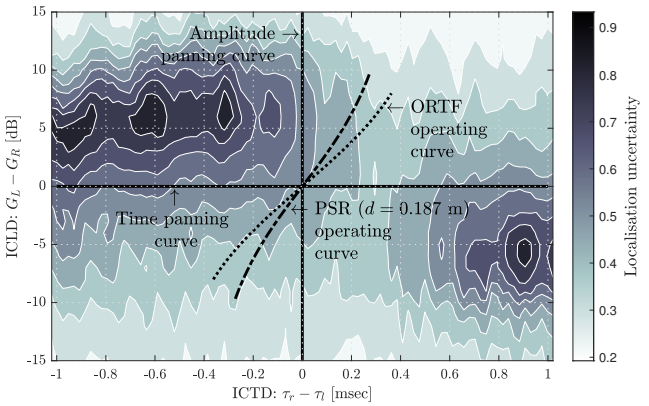
Notice how amplitude panning methods, which are associated to the line ICTD $= 0$ ms (corresponding to the y-axis), lie in an area with increased localization uncertainty for these off-center positions. For the rightward positions in Fig. 10b and 10c, the localization uncertainty is particularly high for positive ICLD values, which are meant to render phantom sources between the midline and the left loudspeaker.

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

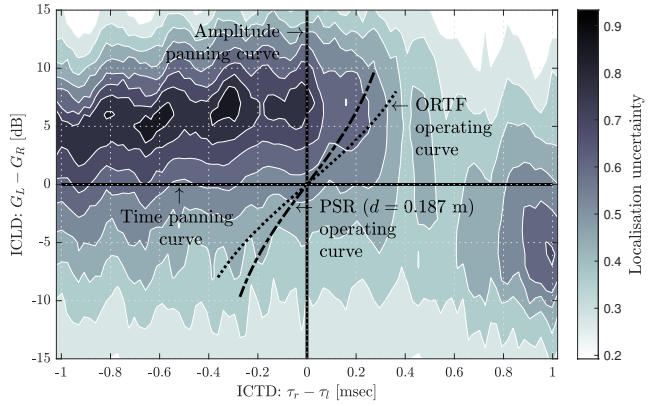
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing



(a) Left off-center position $\mathbf{x} = \{-10, 0\}$ cm



(b) Right off-center position $\mathbf{x} = \{10, 0\}$ cm



(c) Right off-center position $\mathbf{x} = \{20, 0\}$ cm

Fig. 10. Contour plot of localization uncertainty as a function of ICLD and ICTD for a listener in the off-center positions $\mathbf{x} = \{-10, 0\}$ cm, $\mathbf{x} = \{10, 0\}$ cm and $\mathbf{x} = \{10, 0\}$ cm. Overlaid are the time-amplitude panning curve associated to the PSR method with $d = 18.7$ cm and to the ORTF microphone pair.

Viceversa, for the leftward position Fig. 10a, the localization uncertainty is high for negative ICLD values, which are meant to render phantom sources between the midline and the right loudspeaker. Notice also how the PSR panning curve associated to $d = 18.7$ cm largely avoids areas with higher localization uncertainty in both leftward and rightward positions 10 cm off-center.

C. Definition of relative inter-channel time and level differences

In order to aid the interpretation of Figures 10, it is useful to define two new quantities: the *relative* ICTD and ICLD, i.e. equivalent inter-channel time and level difference as they are observed at a point away from the center of the loudspeaker array. These quantities will be referred to as relative inter-channel time difference (RICTD) and relative inter-channel level difference (RICLD), respectively.

Let the observation point be positioned at (x, y) (see Fig. 1). Appendix B provides a closed-form approximation for the RICLD and RICTD under the assumption of small displacement compared to the loudspeaker distance:

$$\begin{aligned} \text{RICLD} &\approx \text{ICLD} - \frac{x}{r_l} \frac{20 \sin\left(\frac{\phi_0}{2}\right)}{\log_e(10)} \ , \\ \text{RICTD} &\approx \text{ICTD} - x \frac{2}{c} \sin\left(\frac{\phi_0}{2}\right) \ . \end{aligned} \tag{19}$$

D. Considerations on the effect of RICTD and RICLD in amplitude panning

Consider now again the $\mathbf{x} = \{10, 0\}$ cm off-center plot in Fig. 10b. The point ICTD $= 0$ ms and ICLD $= 5$ dB (i.e. the left loudspeaker leads in amplitude) is associated to RICTD $= -0.29$ ms and RICLD $= 4.7829$ dB, which have opposite sign and thus provide inconsistent information. The RICLD is now smaller than ICLD $= 5$ dB because the right loudspeaker is closer to the listener. The change, however, is small (0.22 dB), and the left loudspeaker still leads in amplitude. The RICTD, on the other hand, has changed significantly with respect to ICTD, and the right loudspeaker is now leading in time. To summarise, the left loudspeaker is louder and thus leads in terms of amplitude, but the right loudspeaker is now closer and thus leads in terms of time. Even though the ICTD and ICLD were consistent, the horizontal shift of the listener caused the RICTD and RICLD to become inconsistent, which, in turn, led to a high localization uncertainty.

Negative values of ICLD, on the other hand, are less problematic in terms of localization uncertainty for a listener in this position. For instance, ICTD $= 0$ ms and ICLD $= -5$ dB is associated to relative values RICTD $= -0.29$ ms and RICLD $= -5.2171$ dB, which have the same sign and provide consistent information.

Similar arguments can be made for a listener moving to the left (i.e. $x < 0$), as shown in Fig. 10a. In this case, however, the plots shift horizontally to the *left* and the critical area of the plot becomes the bottom one.

In summary, amplitude methods lead to higher localization uncertainty whenever the listener moves to one side but aims to render sound sources in directions closer to the opposite side. Section IV will show that time-amplitude methods provide a lower localization uncertainty in positions away from the center of the sweet-spot.

E. Further considerations on RICTD and RICLD

Notice how the RICLD is a function of $\frac{x}{r_l}$, while the RICTD is independent from r_l. This implies that for large

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

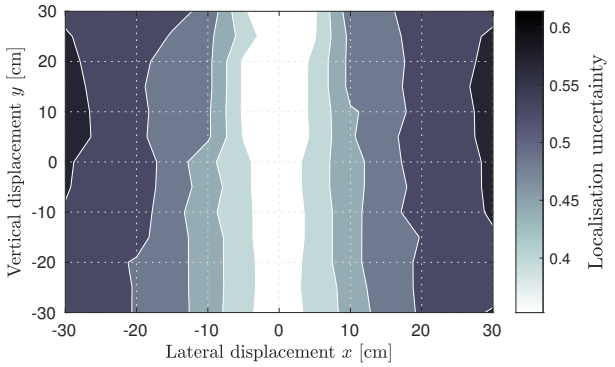


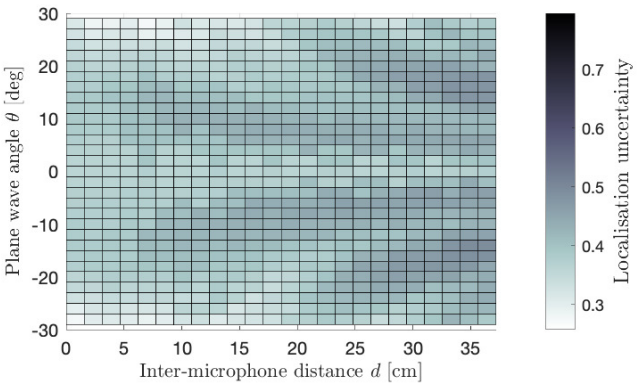
Fig. 11. Localization accuracy of PSR for $d = 0$ cm (i.e. coincident PSR) as a function of the listener position around the center of the sweet spot.

loudspeaker arrays, movement of the observation point away from the center of the sweet-spot causes a larger change of RICTD compared to RICLD. This holds also even for living-room-sized arrays. Consider for instance the RICTD that will result in full perceived shift of the phantom source to one loudspeaker, i.e. RICTD $= 1$ ms. Assuming ICTD $= 0$ ms, ICLD $= 0$ dB and $\phi_0 = \pi/3$ (60 degrees), this is already achieved at $x = \frac{0.001c}{2\sin(\phi_0/2)} = 0.34$ m. In that position, the RICLD, on the other hand, is only 1.5 dB. In other words, a phantom source that appears halfway between loudspeakers for a listener in the center of the sweet-spot, collapses onto one of the loudspeakers when the listener moves to a position 34 cm on either side. This compares to approximately 3 cm on either side for binaural cross-talk cancellation using two loudspeakers [53]. It should be noted that the degradation is gradual in both cases, which makes defining quantitatively the sweet-spot size difficult.

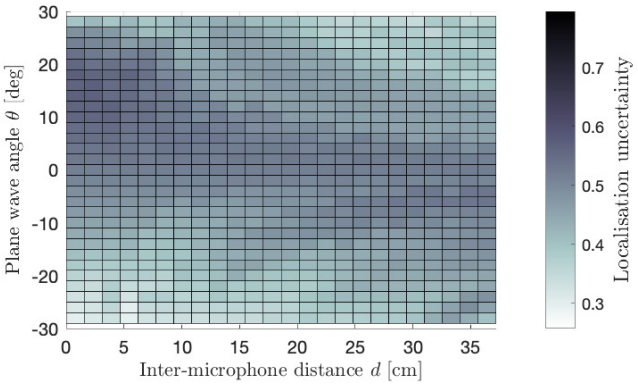
Vertical displacement is less significant than horizontal displacement. This can be inferred by the fact that the first-order approximation (19) does not depend on y (it can be shown that y appears starting from the second-order term). This is confirmed by Fig. 11, which shows the localization uncertainty as a function of listener position for PSR panning/recording associated to $d = 0$ cm. The plot shows that the localization uncertainty varies almost exclusively as a function of the lateral displacement, x. For this reason, the simulations in the remainder of the paper will focus on listener displacement along the x axis ($y = 0$).

V. Time-Amplitude Panning/Recording Methods

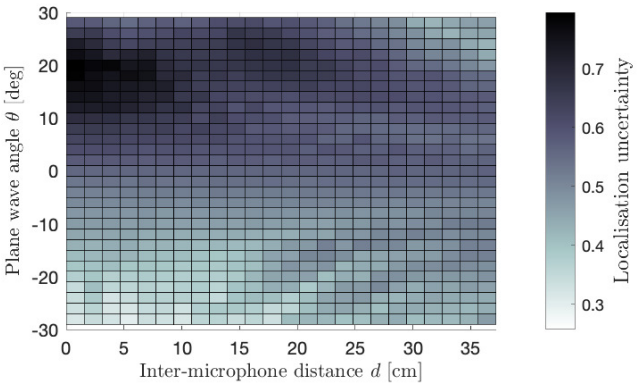
This section will show how small ICTDs allow to render consistent RICLD-RICTD values (and thus a lower localization uncertainty) in a slightly larger area around the center of the sweet-spot. The starting point of the analysis will be PSR panning/recording, which enables to vary the time-amplitude trade-off using the the inter-microphone distance, d, as a free parameter. Then, the performance of popular microphone arrangements will be assessed. The section concludes with some comments on how the results extend to the multichannel case.



(a) On-center



(b) Off-center by 10 cm



(c) Off-center by 20 cm

Fig. 12. Localization uncertainty as a function of inter-microphone distance, d, and plane wave angle, θ_s, in (a) the center position, (b) in a position 10 cm off-center to the right, i.e. $(x, y) = (10, 0)$ cm, and (c) in a position 20 cm off-center to the right, i.e. $(x, y) = (20, 0)$ cm. The colorbar is the same for all three figures to facilitate a comparison between the three figures.

A. Effect of time-amplitude trading on localization uncertainty using PSR panning/recording

Fig. 12 shows the localization uncertainty for PSR panning/recording as a function of inter-microphone distance d and plane wave angle θ_s for three different observation positions. A number of observations can be made:

- As the distance from the center of the sweet spot increases, so does the maximum localization uncertainty.
- In the center position (Fig. 12a), the localization uncertainty is lowest around $d = 0$ (i.e. amplitude methods). Two

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

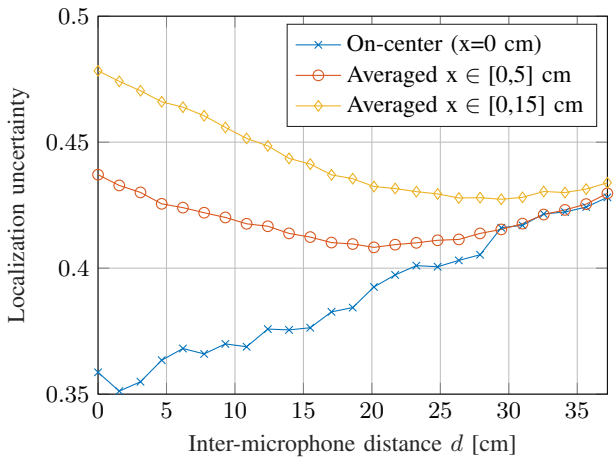


Fig. 13. Localization uncertainty as a function of inter-microphone distance d. Results are averaged across angles between the midline and the left loudspeaker.

"sidebands" appear in the plot beyond $d = 20$ cm. These sidebands correspond to the vertical stripes observed in Fig. 9 at $\pm\tau_o$.

- At $x = 10$ cm (Fig. 12b), the localization uncertainty is higher for inter-microphone distance $d = 0$ cm, i.e. amplitude methods. It remains high in plane wave directions around the midline, while it reduces for other directions, especially around the $\theta_s \in [10°, 20°]$ range. This can be observed also in Fig. 10b, which shows the panning curve associated to $d = 18.7$ cm. Here, the panning curves avoids the areas with higher localization uncertainty coming in from the left.
- At $x = 20$ cm (Fig. 12c), the localization uncertainty tends to reduce for higher d, where it also has a lower variability as a function of source angles θ_s.

Fig. 13 shows the localization uncertainty as a function of inter-microphone distance d. Here the results are averaged across angles between the midline and the left loudspeaker, which are more difficult to render when the listener moves to the right of the center. The curve associated to the center position shows that, in that position, the inter-microphone distance with the lowest uncertainty is close to $d = 0$ cm, as expected. In typical use cases, one would like to minimize the uncertainty as it is observed in a whole region around the center of the sweet spot, instead of specific positions. Fig. 13 shows the localization uncertainty averaged for listener displacements between $x = 0$ cm and $x = 5$ cm (with a 1 cm resolution). Here, the absolute minimum is achieved at $d = 20$ cm. Fig. 13 also shows the localization uncertainty averaged for displacements between $x = 0$ cm and $x = 15$ cm. In this case, the average uncertainty increases significantly at small d, and the absolute minimum moved to $d \approx 30$ cm.

In summary, the results suggest that smaller inter-microphone distances (amplitude methods) are preferable for positions close to the center, while larger inter-microphone distances result in lower localization uncertainty at larger distances. There is thus a trade off between lower localization uncertainty in the center of the sweet spot and away from the center.

Notice that if one can assume that only one listener is present, and that some form of user tracking is available, then the results here suggest that it would be beneficial to correct the ICTDs such that the RICTD $= 0$. In other words, if the location of the (single) listener is known, one can correct the time delays such that the listener experiences amplitude panning.

If, on the other hand, the listener's position is not being tracked, it is typically difficult to predict how far the listener/s will move away from the center of the sweet spot. A possible compromise is to use the inter-microphone distance that results in the largest ICTD to be no larger than τ_o, and thus avoid the areas with slightly higher localization uncertainty in the first and third quadrant shown in Fig. 9. One can find this inter-microphone distance by equating ICTD_{max} with τ_o, and isolating d:

$$d = r_h \frac{\cos\left(\theta_e - \frac{\phi_0}{2}\right) + \frac{\phi_0}{2} + \theta_e - \frac{\pi}{2}}{\sin\left(\frac{\phi_0}{2}\right)} \,. \tag{20}$$

For instance, the inter-microphone distance for $\theta_e = 100°$ and $r_h = 9$ cm, is $d = 18.7$ cm, which is the one used for the panning curve shown in Fig. 9 and 10.

B. Comparison with popular stereophonic microphone arrangements

This subsection presents the results of a comparison between three versions of PSR for $d = 0, 18.7, 37.2$ cm, and the popular stereophonic microphone arrangements summarized in Table I. The comparison involves choosing which source angles θ_s to simulate. While it is clear that for PSR the source angles should be within the microphone base-angle ($60°$), the choice for other microphone arrangements is ambiguous. Indeed, sound engineers typically make a choice of the relative position of the sources of interest (e.g. an orchestra) often based on artistic considerations. In order to make a more unambiguous choice here, the range of angles is chosen according to which source angles results in ICTD-ICLD pairs within the Williams curves. The resulting ranges are termed here *coverage angles*, and are reported in Table I for the popular microphone arrangements. In the simulations that follow, a set of thirty θ_s are uniformly taken within the coverage angle for each microphone arrangement.

Fig. 14 shows the localization uncertainty as a function of lateral displacement x. Fig. 14a shows the average across source angles, while Fig. 14b shows the excursion across source angles, defined as the difference between the largest and the smallest localization uncertainty across source angles. A number of observations can be made:

- The amplitude methods, i.e. Blumlein pair, PSR with $d = 0$ cm, and cardioid XY pair (all of which are indicated by a $\times$ marker) have very similar performance; the difference between them stems from the slightly different distribution of ICLDs across source angles θ_s.
- In comparison to time-amplitude methods, the average uncertainty of the amplitude methods is lower for $|x| < 2$ cm but higher beyond $|x| > 10$ cm.

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

- The uncertainty excursion of the amplitude methods is significantly higher than time-amplitude methods beyond $|x| > 10$ cm, reaching almost 0.5 at $|x| = 20$ cm. This is associated to the difficulty in rendering sound sources associated to direction opposite to the lateral movement of the listener, as discussed in Section IV-D, and it means that that sound sources at different angles will be perceived differently.
- In the center ($x = 0$ cm), the average uncertainty is lowest for the amplitude methods, and then, in order, for PSR ($d = 18.6$ cm), ORTF ($d = 17$ cm), DIN ($d = 20$ cm), PSR ($d = 37.2$ cm) and NOS ($d = 30$ cm). PSR ($d = 18.6$ cm) is lowest among time-amplitude methods because it is the only one that does not cross the $\pm\tau_o$ threshold (see Fig. 9).
- The PSR with $d = 18.6$ cm has a reasonably low average uncertainty across the lateral displacement, a low excursion within $|x| < 15$ cm, but a higher excursion compared to other time-amplitude methods beyond $|x| > 15$ cm.
- The NOS and PSR version with $d = 37.2$ cm have a higher average uncertainty in the center, but a smaller variation of the average uncertainty across x, and a low overall excursion.
- The PSR ($d = 18.6$ cm), ORTF ($d = 17$ cm), DIN ($d = 20$ cm), and NOS ($d = 30$ cm) all have low excursion, with little discernible difference between them across x.

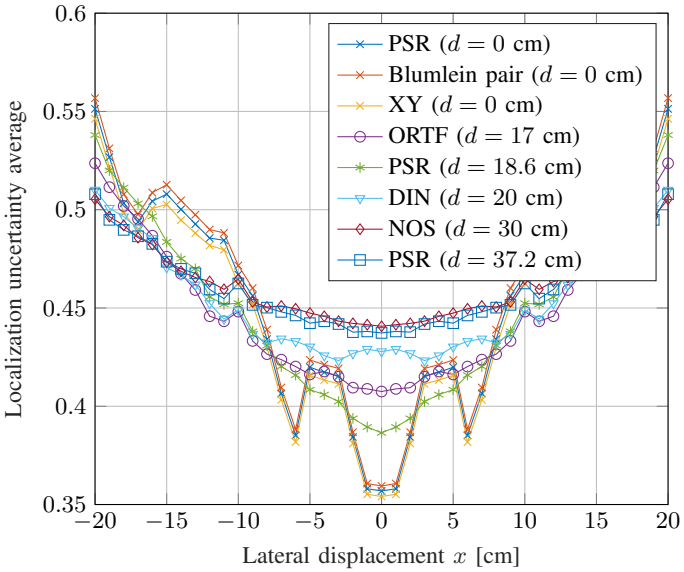
Overall, these results further confirm that amplitude methods have a better performance at the center of the sweet-spot, while time-amplitude methods have a lower overall uncertainty and a lower excursion away from the sweet-spot. A lower excursion means that whenever multiple sound sources are present at different angles, they are rendered with a similar localization uncertainty, which is an appealing property.

A further advantage of PSR over the other time-amplitude methods, is that it has been shown to have a high localization *accuracy*, meaning that subjects would perceive sound sources in directions close to the intended θ_s [21]. This is also related to the fact that in PSR the coverage angle is identical to the loudspeaker base-angle, while the other methods will have (at least) a compression/decompression of angles from the coverage angle to the loudspeaker base-angle. An additional advantage of having the coverage angle identical to the loudspeaker base-angle (and microphone base-angle) is that the extension to the multichannel case is straightforward.

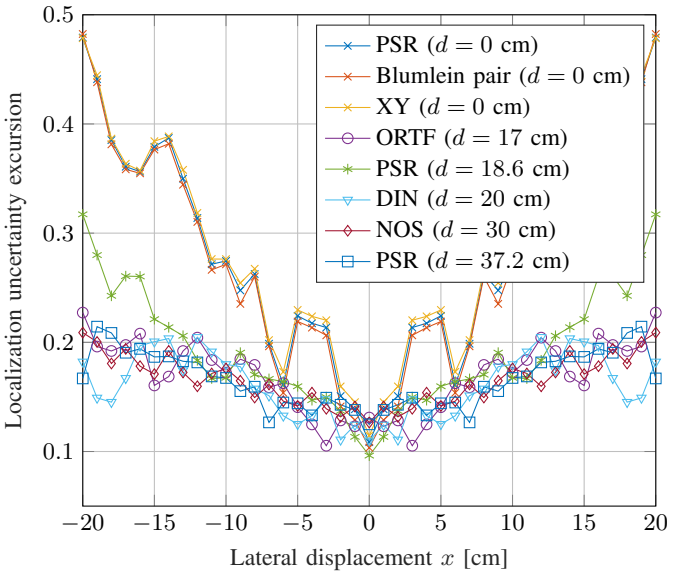
C. Comments on the relation to multichannel PSR

The PSR stereophonic setup considered thus far can be seen as a subset of a multichannel circular array. In [21], it was shown that having only the two loudspeakers closest to θ_s as active yields a lower spatial fluctuation of the active intensity vector field and thus a larger sweet spot. This turns the overall multichannel design problem into a number of stereophonic problems.

Consider the case of M microphones distributed uniformly around the circle with $\phi_m = \phi_0 = 2\pi/M$. Each of these microphones is connected to a loudspeaker in the same angular configuration, without mixing. Johnston and Lam proposed in [22] a microphone array with $M = 5$, resulting in a base angle of $\phi_0 = 2\pi/5 = 72°$, and radius $r_m = 15.5$ cm. This array radius is very close to the value needed to avoid ICTD larger than $\tau_o = \pm 0.3$ ms. Indeed, using (1) and (20) for $\phi_0 = 2\pi/5 = 72°$ yields $r_m = 16.2$ cm. Johnston and Lam stated that $r_m = 15.5$ cm would conserve ITD cues that the listener would have experienced in the recording space, but without providing a rigorous explanation (it could indeed be argued, as Bernfeld did in [48], that correct ITD cues can also be achieved by using ICLDs alone). This paper shows that Johnston and Lam's choice of array radius is the one that leads to a good compromise between low localization uncertainty in central and non-central listener positions. The same choice was made in [21], where the directional accuracy was also improved significantly.



(a) Localization uncertainty averaged across θ_s angles.



(b) Localization uncertainty excursion across θ_s angles calculated as the Difference between maximum and minimum localization uncertainty.

Fig. 14. Comparison of different PSR versions with popular stereophonic microphone arrangements.

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

VI. Summary and Conclusions

This paper focused on the effect of inter-channel time and level differences on perceived localization uncertainty. A computational model was proposed, based on calculating a distance functional between the observed ILD-ITD values and the ones associated to free-field sound sources. The distance functional was chosen as the 0.7-norm which is capable of modelling the splitting of auditory events observed experimentally in case of contradicting ILD-ITD cues. The model predictions had a high correlation with results of formal listening experiments. The model can also be used to predict the perceived localization angle, but this is left to future research.

The model was then used to predict the localization uncertainty under stereophonic reproduction. It was found that contradicting ICTD-ICLD pairs were associated to a high localization uncertainty. Closed form approximations of the ICTDs and ICLDs relative to the listener (denoted as RICTD and RICLD) were presented. It was observed that when a listener moves away from the center of the sweet-spot, the RICLD remains largely unchanged, while the RICTD changes even for small displacements. In off-center positions, then, one can obtain contradicting RICLD-RICTD pairs even if the original ICLD-ICTD pairs were not. It was then explained that non-zero ICTDs (i.e. time-amplitude methods) will delay the onset of contradicting RICLD and RICTD. More significant use of ICTDs results in a higher uncertainty in the center of the sweet-spot, but a relatively lower one in off-center positions. A comparison of amplitude methods (Blumlein, XY) and time-amplitude methods (PSR, ORTF, DIN and NOS) confirms that time-amplitude methods have a lower localization uncertainty in off-center positions, and that the localization uncertainty has a lower variability as a function of source angle.

These results reveal that near-coincident microphone pairs, which are often used for their sense of spaciousness deriving from lower inter-channel decorrelation [6], also have an advantage in terms of localization uncertainty.

In this paper it was not possible to analyze widely spaced microphone arrays such as the ones criticized by Lipshitz [12] or time panning as criticized by Lee and Rumsey [11], since they exceed the 1 ms ICTD limit beyond which one has to also model the law of the first front. This paper is therefore not in contradiction with the findings in [12] and [11], as the conclusions made here are limited to coincident and near-coincident microphone arrays. The computational model could be modified to account for the law of the first front, e.g. by means of inhibitions mechanisms [34], thus allowing analysis of widely spaced microphone arrays.

Throughout this paper, it was implicitly assumed that a low localization uncertainty is desirable. This is motivated by the fact that an actual plane wave is well localized spatially, and that to render it accurately it needs to have a low localization uncertainty. This may not be necessarily the case e.g. in an artistic context, where one may prefer to render sound sources as difficult to localize. The results in this paper are useful in this context too as they can also be read with an implicit preference for high uncertainty.

Appendix A

This appendix provides a closed-form approximation of the ICTD that results in the two loudspeaker signals to arrive at the same time at the left ear (the corresponding case for the right ear will follow). Let the head radius be denoted by r_h and the angle between the forward-looking direction and the ear be denoted by θ_e. Let $\theta_0 = \cos^{-1}\left(\frac{r_h}{r_l}\right)$ be the angle of tangential incidence on the spherical head. Assuming that $\left|\frac{r_h}{r_l}\right| << 1$, i.e. the head radius is much smaller than the loudspeaker distance, then $\theta_0 \approx \frac{\pi}{2}$.

The angle between the left loudspeaker and the left ear is $\theta_{LL} = \theta_e - \frac{\phi_0}{2}$ and under most conditions, $\theta_{LL} < \theta_0$. The distance between the left loudspeaker and the left ear is then:

$$d_{LL} = \sqrt{r_l^2 + r_h^2 - 2 r r_h \cos(\theta_{LL})} \tag{21}$$

$$= r_l \sqrt{1 + \frac{r_h^2}{r_l^2} - 2\frac{r_h}{r_l}\cos\left(\theta_e - \frac{\phi_0}{2}\right)} \tag{22}$$

$$\approx r_l - r_h \cos\left(\theta_e - \frac{\phi_0}{2}\right) \ , \tag{23}$$

where in the last step, the quadratic term $\frac{r_h^2}{r_l^2}$ is ignored and the square root is approximated using the first-order Taylor series approximation.

The angle between the right loudspeaker and the left ear is $\theta_{RL} = \frac{\phi_0}{2} + \theta_e$ and under most conditions, $\theta_{RL} > \theta_0$. The distance between the right loudspeaker and the left ear, then, is given by the summation of the distance between the loudspeaker and the point of tangential incidence, and the remaining angular sector:

$$d_{RL} = \sqrt{r_l^2 + r_h^2} + r_h(\theta_{RL} - \theta_0) \ . \tag{24}$$

For small $\frac{r_h}{r_l}$, the quadratic term $(\frac{r_h}{r_l})^2$ can be ignored:

$$d_{RL} \approx r_l + r_h\left(\frac{\phi_0}{2} + \theta_e - \frac{\pi}{2}\right) \ . \tag{25}$$

The delay applied to the left loudspeaker, τ_o, that makes the two loudspeaker signals arrive at the same time at the left ear satisfies the equation

$$\frac{d_{LL}}{c} + \tau_o = \frac{d_{RL}}{c} \ . \tag{26}$$

Replacing equations (23) and (25), and solving for τ_o yields

$$\tau_o \approx \frac{r_h}{c}\left[\cos\left(\theta_e - \frac{\phi_0}{2}\right) + \frac{\phi_0}{2} + \theta_e - \frac{\pi}{2}\right] \ . \tag{27}$$

Due to the symmetry of the problem, the ICTD that results in the loudspeaker signals arriving at the same time at the *right* ear is $-\tau_o$.

Appendix B

This appendix provides a closed-form approximation of the RICTD and RICLD. Towards this end, approximations of the relative distance between an observation point (x, y) and the two loudspeakers at a distance r_l from the center of the sweet-spot $(0, 0)$ are sought first. It will be assumed that $\left|\frac{x}{r_l}\right| <<$

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

1 and $\left|\frac{y}{r_l}\right| << 1$. The distance between the observation point (x, y) and the right loudspeaker can be approximated as

$$d_R^2 = \left(r_l \sin\left(\frac{\phi_0}{2}\right) - x\right)^2 + \left(r_l \cos\left(\frac{\phi_0}{2}\right) - y\right)^2 \quad (28)$$

$$\approx r_l^2 \left[1 - 2\frac{x}{r_l}\sin\left(\frac{\phi_0}{2}\right) - 2\frac{y}{r_l}\cos\left(\frac{\phi_0}{2}\right)\right] , \quad (29)$$

where in the second step it is assumed that $\left(\frac{x}{r_l}\right)^2 \approx 0$, and $\left(\frac{y}{r_l}\right)^2 \approx 0$. It is useful to define $a = 2\frac{x}{r_l}\sin\left(\frac{\phi_0}{2}\right)$ and $b = 2\frac{y}{r_l}\cos\left(\frac{\phi_0}{2}\right)$, so that d_R can be rewritten as $d_R \approx r_l\sqrt{1-(a+b)}$. The first-order Taylor series approximation of the square root yields:

$$d_R \approx r_l\left(1 - \frac{a+b}{2}\right) \quad (30)$$

Using similar steps, d_L can be approximated by $d_L \approx r_l\left(1 + \frac{a-b}{2}\right)$. The time delay between the two loudspeakers observed in (x, y) can now be approximated as

$$\frac{d_R - d_L}{c} \approx -\frac{r_l a}{c} = -x\frac{2}{c}\sin\left(\frac{\phi_0}{2}\right) . \quad (31)$$

The relative level observed in (x, y) for the two point-like loudspeakers can now be approximated as

$$20\log_{10}\left(\frac{\frac{1}{d_L}}{\frac{1}{d_R}}\right) \approx 20\log_{10}\frac{1 - \frac{x}{2r_l}\sin\left(\frac{\phi_0}{2}\right) - \frac{y}{2r_l}\cos\left(\frac{\phi_0}{2}\right)}{1 + \frac{x}{2r_l}\sin\left(\frac{\phi_0}{2}\right) - \frac{y}{2r_l}\cos\left(\frac{\phi_0}{2}\right)} , \quad (32)$$

and, using a first-order Taylor series approximation:

$$20\log_{10}\left(\frac{\frac{1}{d_L}}{\frac{1}{d_R}}\right) \approx -\frac{x}{r_l}\frac{20}{\log_e(10)}\sin\left(\frac{\phi_0}{2}\right) , \quad (33)$$

Using (31) and (33), one obtains RICLD $\approx$ ICLD $- \frac{x}{r_l}$ and RICTD $\approx$ ICTD $- x\frac{2}{c}\sin\left(\frac{\phi_0}{2}\right)$.

ACKNOWLEDGMENTS

The authors would like to thank Niccolò Antonello for pointing out that the proposed computational model has a statistical interpretation.

REFERENCES

[1] E. De Sena and Z. Cvetković, "A computational model for the estimation of localisation uncertainty," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process. (ICASSP-13)*, Vancouver, Canada, May 2013, pp. 388–392.

[2] E. De Sena, "Analysis, design and implementation of multichannel audio systems," Ph.D. dissertation, King's College London, 2013.

[3] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Design of a circular microphone array for panoramic audio recording and reproduction: Array radius," presented at the AES 128^{th} Conv., Preprint #8064, London, UK, May 2010.

[4] H. Hacıhabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. Smith, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36–54, 2017.

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

[6] F. Rumsey, *Spatial audio*. Focal Press, 2001.

[7] V. Pulkki, "Virtual sound source positioning using vector-base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[8] J. Eargle, *The microphone book*. Focal Press, 2004.

[9] J. Daniel, J. Rault, and J. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," presented at the 105th Audio Eng. Soc. Conv., Preprint #4795, San Francisco, CA, USA, Sep. 1998.

[10] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28, 2013.

[11] H. Lee and F. Rumsey, "Level and time panning of phantom images for musical sources," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 978–988, 2013.

[12] S. P. Lipshitz, "Stereo microphone techniques: Are the purists wrong?" *J. Audio Eng. Soc*, vol. 34, no. 9, pp. 716–744, May 1986.

[13] F. Rumsey and T. McCormick, *Sound and recording: applications and theory*. Focal Press, 2014.

[14] M. Plewa and P. Kleczkowski, "Choosing and configuring a stereo microphone technique based on localisation curves," *Archives of Acoustics*, vol. 36, no. 2, pp. 347–363, 2011.

[15] H. Wittek and G. Theile, "Development and application of a stereophonic multichannel recording technique for 3d audio and vr," in *presented at the 143rd Audio Eng. Soc. Conv., Preprint #9869, New York, USA*. Audio Engineering Society, 2017.

[16] L. Riitano, M. Victoria, and J. Enrique, "Comparison between different microphone arrays for 3d-audio," in *presented at the 144th Audio Eng. Soc. Conv., Preprint #9980, Milan, Italy*. Audio Engineering Society, 2018.

[17] M. Williams, *Microphone Array Analysis for Stereo and Multichannel Sound Recording*. Editrice Il Rostro, 2004, vol. 1.

[18] H. Lee, "Capturing 360 audio using an equal segment microphone array (esma)," *J. Audio Eng. Soc.*, vol. 67, no. 1/2, pp. 13–26, 2019.

[19] C. Millns and H. Lee, "An investigation into spatial attributes of 360° microphone techniques for virtual reality," in *presented at the 144th Audio Eng. Soc. Conv.*, 2018.

[20] H. Lee, D. Johnson, and M. Mironovs, "An interactive and intelligent tool for microphone array design," in *presented at the 143rd Audio Eng. Soc. Conv., e-Brief #390, New York, USA*, 2017.

[21] 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.*, vol. 21, no. 8, pp. 1653–1665, Aug. 2013.

[22] J. D. Johnston and Y. H. Lam, "Perceptual soundfield reconstruction," presented at the 109th Audio Eng. Soc. Conv., Preprint #2399, Los Angeles, CA, USA, Sep. 2000.

[23] H. Hacıhabiboğlu and Z. Cvetković, "Panoramic recording and reproduction of multichannel audio using a circular microphone array," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA'09)*, oct. 2009, pp. 117 –120.

[24] G. W. Elko, "Differential microphone arrays," in *Audio signal processing for next-generation multimedia communication systems*, Y. Huang and J. Benesty, Eds. Kluwer Academic Publishers, 2004.

[25] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "On the design and implementation of higher order differential microphones," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 1, pp. 162–174, Jan. 2012.

[26] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. John Wiley & Sons, 2006.

[27] 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.

[28] M. B. Gardner, "Historical background of the haas and/or precedence effect," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1243–1248, 1968.

[29] M. Williams and G. Le Du, "Microphone array analysis for multichannel sound recording," presented at the 107th Audio Eng. Soc. Conv., e-Brief #390, New York, USA, Sep. 1999.

[30] N. V. Franssen, *Stereophony*. Philips Research Laboratories, 1964.

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

[32] H. Hacıhabiboğlu, E. De Sena, and Z. Cvetković, "Design of a circular microphone array for panoramic audio recording and reproduction: Microphone directivity," presented at the 128th Audio Eng. Soc. Conv., Preprint #8063, London, UK, May 2010.

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

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2975419, IEEE/ACM Transactions on Audio, Speech, and Language Processing

[34] 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. 1608–1622, Jul. 1986.
[35] 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.
[36] 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, Jan. 2005.
[37] 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, p. 3075, Jul. 2004.
[38] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. of Signal Proc. to Audio and Acoust. (WASPAA'01)*, New Paltz, NY, USA, 2001, pp. 99–102.
[39] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple, Perception Group, Tech. Rep.*, 1993.
[40] 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, Aug. 1990.
[41] A. N. Kolmogorov and S. V. Fomin, *Elements of the theory of functions and functional analysis*. Dover Publications, 1999.
[42] B. Supper, "An onset-guided spatial analyser for binaural audio," Ph.D. dissertation, University of Surrey, 2005.
[43] S. S. Stevens, "On the psychophysical law," *Psychological review*, vol. 64, no. 3, pp. 153–181, 1957.
[44] I. R. Goodman and S. Kotz, "Multivariate θ-generalized normal distributions," *J. of Multivariate Analysis*, vol. 3, no. 2, pp. 204–219, 1973.
[45] E. De Sena, M. Brookes, P. A. Naylor, and T. v. Waterschoot, "Localization experiments with reporting by head orientation: statistical framework and case study," *J. Audio Eng. Soc.*, vol. 65, no. 12, pp. 982–996, 2017.
[46] K. V. Mardia and P. E. Jupp, *Directional statistics*. Hoboken, New Jersey: John Wiley & Sons, 2009.
[47] Various, *Recomm. BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems*. ITU-R, 2003.
[48] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," presented at the 44th Audio Eng. Soc. Conv., Preprint #C-4, Rotterdam, the Netherlands, Mar. 1973.
[49] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Proc. 23rd Audio Eng. Soc. Int. Conf.*, Copenhagen, Denmark, May 2003.
[50] M. Poletti, "A unified theory of horizontal holographic sound systems," *J. Audio Eng. Soc.*, vol. 48, no. 12, pp. 1155–1182, Dec. 2000.
[51] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Univ. of Paris VI, France, July 2000.
[52] D. M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *J. Acoust. Soc. Am.*, vol. 31, no. 7, pp. 977–986, 1959.
[53] J. Rose, P. Nelson, B. Rafaely, and T. Takeuchi, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 1992–2002, 2002.

2329-9290 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:52:07 UTC from IEEE Xplore. Restrictions apply.