Towards Maximizing a Perceptual *Sweet Spot* for Spatial Sound with Loudspeakers

Pedro Izquierdo Lehmann, *Student Member, IEEE,* Rodrigo F. Cádiz, *Member, IEEE,* and Carlos A. Sing Long, *Member, IEEE*

***Abstract*—The *sweet spot* can be interpreted as the region where acoustic sources create a spatial auditory illusion. We study the problem of maximizing this sweet spot when reproducing a desired sound wave using an array of loudspeakers. To achieve this, we introduce a theoretical framework for spatial sound perception that can be used to define a sweet spot, and we develop a method that aims to generate a sound wave that directly maximizes the sweet spot defined by a model within this framework. Our method aims to incorporate perceptual principles from the onset and it is flexible: it imposes little to no constraints on the regions of interest, the arrangement of loudspeakers or their radiation pattern. However, the perceptual models must satisfy a convexity condition, which is fulfilled by state-of-the-art monaural perceptual models, but not by binaural ones. Proof-of-concept experiments show that our method, when implemented with van de Par's monaural model, outperforms state-of-the-art sound field synthesis methods in terms of their binaural azimuth localization and binaural coloration properties.**

***Index Terms*—Spatial sound, sound field synthesis, sweet spot, perception, psycho-acoustics, non-convex optimization.**

I. Introduction

The field of spatial sound addresses the question: *how do we create a desired* spatial auditory illusion *over a spatial region of interest with a set of acoustic sources?* [1, Chapter 2.3]. The *spatial auditory illusion* (SAI) occurs when acoustic sources create a sound scene that produces a desired auditory scene over a region. It is related to *sound quality* as described by Wierstof et al [2, Chapter 2]: *"The quality of a system as perceived by a listener is considered to be the result of assessing perceived features with regard to the desired features (of the auditory scene)."* Following Blauert, the *sound scene* represents the objective nature of a sound wave propagating in the physical world, whereas the *auditory scene* represents the imprint of the sound scene in our subjectivity, that is, the result of the auditory system perceiving and organizing sound into meaning [3]. Over the last century, several methods have been proposed to answer this question. Their performance can be compared in terms of the size of the region where the SAI is achieved. In this work, we call this region the *sweet spot*.

The term *sweet spot* is already used in panning and surround systems to describe an ideal listening position that is equally distant to all loudspeakers [4] and around which there is a limited area where a desired wavefront is correctly recreated [1, Chapter 3.1], [5]. It is also used to mean the *"area in which the spatial perception of the auditory scene works without considerable impairments"* [6, Chapter 1.2]. Furthermore, in [7] the *sweet area* is defined as the *"area within which the reproduced sound scene is perceived as plausible,"* where *plausible* means the preservation of front localization and envelopment of reverberation. Our use of the term is in the spirit of the second and third ideas.

One of the earlier and most widespread spatial sound approaches is stereophony and its generalizations, surround systems [8] and Vector Base Amplitude Panning (VBAP) [9]. These methods, also called *panning techniques*, adjust the level and time-delay of the audio signals for each speaker utilizing a *panning law* to steer the perceived direction of the sound source. The possibility of this steering has been explained by the perceptual idea of *summing localization* and by the *association model* [1, Chapter 6.1]. Moreover, due to some psycho-acoustic features of the auditory system, such as the binaural decoloration mechanism, they work sufficiently well in some applications, even with few speakers [4]; they do not suffer from coloration [10]. However, they can only simulate sound sources that lie on the segments that join the speakers. Furthermore, its quality degrades rapidly as the listener moves away from the center of the target region [4].

A popular strategy to recreate an auditory scene is to directly approximate the sound wave that created it. In the literature, this strategy is called *sound field synthesis*, *sound field reproduction* or *sound field reconstruction*. Following Huygens' principle [11], any sound scene can be approximated accurately with a sufficiently dense arrangement of loudspeakers. However, in practice there is only a limited number of them. Three classes of commonly used methods for sound field synthesis are *mode matching methods*, *pressure matching methods* and *wave field synthesis*.

Mode Matching Methods (MMM) match the coefficients in the expansion of the target and generated sound waves in spatial spherical harmonics [12]. Some well-known MMMs are Ambisonics [13], Higher-Order Ambisonics (HOA), and Near-Field Compensated Ambisonics (NFC-HOA) [14]. All of them are designed for circular or spherical regions of interest. When approximating a plane wave, they create a central

P. Izquierdo Lehmann and C. A. Sing Long are with the Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile, Chile. E-mail: pizquierdo2@uc.cl and casinglo@uc.cl.

R. F. Cádiz is with the Department of Electrical Engineering, School of Engineering, and Music Institute, Faculty of Arts, Pontificia Universidad Católica de Chile, Chile. E-mail: rcadiz@uc.cl

C. A. Sing Long is also with the Institute of Biological and Medical Engineering, Pontificia Universidad Católica de Chile, Chile, with ANID – Millennium Science Initiative Program – Millennium Nucleus Center for the Discovery of Structures in Complex Data, and with ANID – Millennium Science Initiative Program – Millennium Nucleus Center for Cardiovascular Magnetic Resonance.

spherical region with a radius that is inversely proportional to the frequency of the source over which the sound scene is reconstructed almost identically [15]. Some variations of these methods consider a weighted mode matching problem to prioritize certain spatial regions [16], a mixed pressure-velocity mode matching problem [17], and an intensity mode matching problem [18].

Instead of using expansions in spatial spherical harmonics, Pressure Matching Methods (PMM) minimize the spatio-temporal L^2-error between the target and generated sound waves [19]. The magnitude of the audio signals are often penalized by their L^p-norm to mitigate the effects of ill-conditioning [20]. Typically the loudspeakers are modeled as monopoles. In most cases, the solution can only be found numerically, and the discretization of the region of interest plays an important role [21], [22].

Finally, Wave Field Synthesis (WFS) leverages the single-layer boundary integral representation of a sound wave over a region of interest [4], [23], [24]. It has been shown that the localization properties of the auditory scene are correctly simulated by WFS and do not depend on the position of the listener over the region of interest [25]. However, this technique suffers from coloration effects due to spatial aliasing artifacts [26].

There is extensive literature analyzing these methods and comparing their performance [14], [27], they become equivalent in the limit of a continuum of loudspeakers, differing only when using a finite number of them [28]. Although they are amenable to mathematical analysis and have computationally efficient implementations, their construction has no natural perceptual justification to produce a large sweet spot. As a consequence, the artifacts introduced by these methods, due to approximation errors, may produce noticeable, and possibly avoidable, perceptual artifacts.

An alternative to better reproduce the auditory scene is to explicitly account for psycho-acoustic and perceptual principles in the reconstruction methods [6]. The first steps in this direction were taken in [29] by proposing a simple model that aims to preserve the spatial properties of the desired auditory scene. A method to reproduce an active intensity field that is largely uniform in space was then proposed in [30]. It is based on an optimization problem yielding audio signals where at most two loudspeakers are active simultaneously. However, it makes the restrictive assumption that the target sound wave is a plane wave, and that the loudspeakers emit plane waves. In [31] the *radiation method* and the *precedence fade* are proposed. The former is equivalent to applying a PMM over a selection of frequencies that are most relevant psycho-acoustically, whereas the latter is a method to overcome the localization problems associated to the *precedence effect* [32]. Finally, in [33] a PMM is extended to account for psycho-acoustic effects by considering the L^2-norm of the differences in pressure convolved in time by a suitable filter.

We believe that there is a gap between methods that aim to directly approximate a sound wave to reproduce a desired auditory scene, and methods that leverage perceptual models to reproduce the same auditory scene. Defining the sweet spot requires a model, either theoretical or empirical, of audio perception. In this work, we introduce a theoretical framework for spatial audio perception, and we develop a method to maximize the sweet spot defined by a model within this framework. Our method is amenable to mathematical analysis, has an efficient computational implementation, and is guided by perceptual principles. Our numerical results show that our method outperforms some state-of-the-art methods for sound field synthesis.

The paper is organized as follows. In Section II we introduce the physical assumptions we make, and a theoretical framework for spatial audio perception, deferring to Appendix A the technical details. Then, in Section III we introduce an intuitive and readily implementable instance of our method to maximize the sweet spot defined by a model within this framework. In Section IV we discuss the perceptual concepts that, to our knowledge, can be incorporated in the theoretical framework. In Section V we present an implementation of our method. In Section VI we perform proof-of-concept numerical experiments analyzing the performance of our method, comparing its results with WFS, NFC-HOA and PMM. Finally, in Section VII we discuss our results, the limitations of our method, and some future lines of research.

II. Framework for spatial sound with loudspeakers

A. Acoustic framework

Consider n_s loudspeakers located at positions $x_1, \dots, x_{n_s} \in \mathbb{R}^3$. When the medium is homogeneous and isotropic and each loudspeaker behaves as an isotropic point source, i.e., as a monopole, the physical sound wave u generated is represented in frequency as [34, Section 2.5.2]

$$\widehat{u}(f,x) = \sum_{k=1}^{n_s} \widehat{\alpha}_k(f) \frac{e^{-2\pi i c^{-1} f \|x - x_k\|}}{4\pi \|x - x_k\|} \tag{1}$$

where c is the speed of sound, $\alpha_1, \dots, \alpha_{n_s}$ are the audio signals driving each loudspeaker, and $\widehat{\alpha}_k$ is the Fourier transform of α_k in time

$$\widehat{\alpha}_k(f) := \int \alpha_k(t) e^{-2\pi i f t}\, dt.$$

From now on, we let $\widehat{\ }$ denote the Fourier transform in time. To model the spatial radiation pattern of each loudspeaker, or time-invariant effects such as reverb [35], [36], we may use

$$\widehat{u}(f,x) = \sum_{k=1}^{n_s} \widehat{\alpha}_k(f) G_k(f,x), \tag{2}$$

where G_k is the Green function of the k-th loudspeaker. In addition to the array, we consider a region of interest $\Omega \subset \mathbb{R}^3$ such that $x_k \notin \overline{\Omega}$; thus, it contains no singularity in (1). On this region, we may approximate a sound wave u_0 *as best as possible* with the array of loudspeakers. If we had a continuum of monopoles on $\partial\Omega$ then, under suitable conditions, the *simple source formulation* [37, Section 8.7] shows we can reproduce u_0 *exactly*. However, when only a

finite number of physical loudspeakers are available, we must find $\widehat{\alpha}_1, \dots, \widehat{\alpha}_{n_s}$ such that

$$\widehat{u}_0(f,x) \approx \sum_{k=1}^{n_s} \widehat{\alpha}_k(f) G_k(f,x), \tag{3}$$

in an suitable sense, for $x \in \Omega$. When each G_k is real-analytic on its second argument the approximation cannot be exact on *any* open set unless u_0 was actually generated by the speaker array [38, Corollary 1.2.5]. This suggests that perfect sound field reconstruction is impossible, and that the difference can be small only on average. Even then, the approximation can be *perceptually* accurate in some subset of Ω.

B. Perceptual framework

The comparison in (3) is between two *physical* quantities. To incorporate *perceptual* effects, we formally introduce a theoretical framework and defer the mathematical details to Appendix A-A.

The perception of an individual located at $x \in \Omega$ and looking in the direction represented by a unit vector θ in $\mathbb{R}^3$ (or an angle in $\mathbb{R}^2$) depends on the relation between the sound wave at the left and right ears. Hence, we consider *pairs* of signals u^ℓ, u^r, denoting *left* and *right* signals, so that $u^s = u^s(t,x,\theta)$ represents the wave that reaches the ear $s \in \{\ell, r\}$ of a listener located at x and looking in the direction θ at time t. We let $u^s_{(x,\theta)}$ represent the signal at ear s. We denote $\bar{u} = (u^\ell, u^r)$ this pair of signals, and let W be the space of all such pairs of signals. Then, instead of (2), now we consider

$$\widehat{u}^s(f,x,\theta) = \sum_{k=1}^{n_s} \widehat{\alpha}_k(f) H^s_k(f,x,\theta), \tag{4}$$

where H^s_k is the *head-related transfer function* [3] associating to a wave emitted by the k-th loudspeaker the sound wave reaching the ear s; this comprises the behavior of the loudspeakers. From now on, we let W_S be the set of all pairs of signals generated by model (4) and we always use $\bar{\ }$ to denote pairs of left and right signals.

Therefore, the problem is to approximate the fixed target $\bar{u}_0$ associated to the sound wave u_0 by a $\bar{u}$ where each signal is represented as (4) that is *perceptually close* to $\bar{u}_0$. To model the *perceptual dissimilarity* we introduce a map D that associates to a pair $\bar{u}, \bar{u}_0$ the function $D_{(\bar{u},\bar{u}_0)} = D_{(\bar{u},\bar{u}_0)}(x,\theta)$ that quantifies the dissimilarity between the signals $\bar{u}$ and $\bar{u}_0$ perceived by a listener located at x and looking in the direction θ. We do not make strong assumptions on D *except* that it is *convex* on its first argument: for any choice of pairs of signals $\bar{u}_1, \bar{u}_2$ we must have

$$\begin{aligned} D_{(\lambda\bar{u}_1+(1-\lambda)\bar{u}_2,\bar{u}_0)}(x,\theta) \le{}& \lambda D_{(\bar{u}_1,\bar{u}_0)}(x,\theta) \\ &+ (1-\lambda) D_{(\bar{u}_2,\bar{u}_0)}(x,\theta) \end{aligned} \tag{5}$$

for any $\lambda \in [0,1]$. We assume the dissimilarity is negligible if $D_{(\bar{u},\bar{u}_0)}(x,\theta) \le 0$. Depending on the application, this may be interpreted as *authenticity*, i.e., $\bar{u}$ is indiscernible from $\bar{u}_0$, or as *plausibility*, i.e., some perceived features of $\bar{u}$ and $\bar{u}_0$ show a context-reasonable correspondence [2, Chapter 2]. In Section IV we discuss some functional forms for D.

Suppose a listener at x can look only along some directions of interest Θ_x. We define the *auditory illusion threshold* as

$$T_D\bar{u}(x) := \sup_{\theta\in\Theta_x} D_{(\bar{u},\bar{u}_0)}(x,\theta). \tag{6}$$

A listener located at x will perceive no noticeable differences between $\bar{u}$ and $\bar{u}_0$ regardless of the direction she is looking if $T_D\bar{u}(x) \le 0$. Hence, the *sweet spot*

$$\mathcal{S}(\bar{u}) = \{x \in \Omega : T_D\bar{u}(x) \le 0\} \tag{7}$$

is the region within Ω where a listener does not perceive significant differences between $\bar{u}$ and $\bar{u}_0$. We define a *loudness discomfort threshold* similarly: we assume there is a function L that associates to $\bar{u}$ the function $L_{\bar{u}} = L_{\bar{u}}(x,\theta)$ quantifying the *loudness discomfort* experienced by a listener at a location x looking in a direction θ. We also assume L satisfies (5) and we define the *loudness discomfort threshold*

$$T_L\bar{u}(x) := \sup_{\theta\in\Theta_x} L_{\bar{u}}(x,\theta).$$

We assume that $T_L\bar{u}(x) \le 0$ when no discomfort is experienced. Hence, to avoid choosing a signal $\bar{u}$ that causes discomfort, we restrict our choices to

$$\mathcal{P} := \{\bar{u} \in W : T_L\bar{u}(x) \le 0 \text{ for a.e. } x \in \Omega\}. \tag{8}$$

Consequently, our goal is to find a $\bar{u} \in W_S$ that maximizes the *weighted area* of the sweet spot $\mu(\mathcal{S}(\bar{u}))$ while causing no discomfort by solving

$$(P_0) \quad \underset{\bar{u}\in W_S\cap\mathcal{P}}{\text{maximize}} \quad \mu(\mathcal{S}(\bar{u})).$$

III. The SWEET-ReLU method

We present an instance of our method to approximate the solution to (P_0). We defer the analysis of our general method to Appendix A. Let $h : \mathbb{R} \to \mathbb{R}$ be such that $h(t) = 0$ if $t < 0$ and $h(t) = 1$ if $t \ge 0$. Since

$$\mu(\mathcal{S}(\bar{u})) = \mu(\Omega) - \int_\Omega h(T_D\bar{u}(x))\, d\mu(x),$$

maximizing $\mu(\mathcal{S}(\bar{u}))$ implies minimizing the second term in the right-hand side. This is challenging as h is piecewise constant. Hence, for $\varepsilon > 0$ we define the continuous approximation $\tilde{h}_\varepsilon : \mathbb{R} \to \mathbb{R}$ of h as $\tilde{h}_\varepsilon(t) = 0$ for $t < 0$, $\tilde{h}_\varepsilon(t) = t/\varepsilon$ for $t \in [0,\varepsilon]$ and $\tilde{h}_\varepsilon(t) = 1$ for $t > \varepsilon$. We can solve

$$(P^{\text{SReLU}}_\varepsilon) \quad \underset{\bar{u}\in W_S\cap\mathcal{P}}{\text{minimize}} \quad \int_\Omega \tilde{h}_\varepsilon(T_D\bar{u}(x))\, d\mu(x).$$

Let $\Omega_\varepsilon(\bar{u}) := \{x \in \Omega : T_D\bar{u}(x) \le \varepsilon\}$. The objective can be decomposed as

$$\frac{1}{\varepsilon}\int_{\Omega_\varepsilon(\bar{u})} (T_D\bar{u}(x))_+ d\mu(x) + \mu(\Omega_\varepsilon(\bar{u})^c)$$

where $(t)_+ = \max\{0,t\}$. The first integral is the contribution from the region where we may change $\bar{u}$ to decrease the auditory illusion threshold, whereas the second is the contribution of the region over which it is *already* too large. Thus, we proceed iteratively: choose $\tilde{\Omega}_1 \subset \Omega$, and for $k \ge 1$ define the sets $\tilde{\Omega}_{k+1} = \tilde{\Omega}_k \cap \Omega_\varepsilon(\bar{u}_k)$ where $\bar{u}_k$ is the solution to

$$(P^{\text{SReLU}}_{\varepsilon,k}) \quad \underset{\bar{u}\in W_S\cap\mathcal{P}}{\text{minimize}} \quad \int_{\tilde{\Omega}_k} (T_D\bar{u}(x))_+\, d\mu(x).$$

Algorithm 1: SWEET-ReLU

input: A decreasing sequence $\{\varepsilon_i\}$ of positive numbers, positive integers $n_\varepsilon, n_{\max}$, initial Ω_1

for $i = 1, \dots, n_\varepsilon$ **do**

  for $j = 1, \dots, n_{\max}$ **do**

    $\bar{u}_j \leftarrow \text{SOLVE}(P^{\text{SReLU}}_{\varepsilon,j})$

    $\Omega_{j+1} \leftarrow \Omega_j \cap \{x \in \Omega : T\bar{u}_j(x) \leq \varepsilon_i\}$

  end

  set : $\bar{u}^\star_i \leftarrow \bar{u}_{n_{\max}}$

  set : $\Omega_1 \leftarrow \{x \in \Omega : T_D\bar{u}_{n_{\max}}(x) \leq \varepsilon_i\}$

end

return $\bar{u}^\star_{n_\varepsilon}$

The above problem is convex, whence $\bar{u}_k$ can be found using efficient algorithms. We can show that the sequence $\{\bar{u}_k\}_{k\in\mathbb{N}}$ has at least one accumulation point. We regard this accumulation point $\bar{u}^\star$ as an approximation to the solution to $(P^{\text{SReLU}}_\varepsilon)$ that yields an approximation $\Omega_\varepsilon(\bar{u}^\star)$ to the optimal sweet spot. Once the sequence has converged, we can repeat the procedure for a smaller value $\varepsilon' < \varepsilon$ choosing $\Omega_{\varepsilon'}(\bar{u}^\star)$ as the initial set. By iteratively running the algorithm for increasingly smaller values of ε we expect to obtain an increasingly accurate approximation to a solution to (P_0). Although we cannot prove this at the moment, our results show we obtain reasonable results using this method. We call this instance of our method SWEET-ReLU (Algorithm 1). We defer a justification of these facts to Appendix A and the deduction of this instance to Appendix A-E.

IV. Perceptual and psycho-acoustic theory

The theoretical framework introduced aims to be flexible enough to account for a variety of perceptual and psycho-acoustic models. However, there are some perceptual and psycho-acoustic effects whose present models, to our knowledge, are outside this framework; this does not preclude that these effects may be modeled within our framework in the future. To understand the consequences of this, we now present some of the perceptual and psycho-acoustic considerations that lead to models within this framework.

The *spatial auditory illusion* (SAI) is achieved when the imprint of the reproduced sound scene on a listener resembles a desired auditory scene. The formation of the auditory scene depends not only on the signals reaching the listener's ears but both on the listener itself [39]. Also, it may even be influenced by external visual, tactile, and proprioreceptive stimuli [40]. Formulating a model accounting for all these effects goes beyond the scope of this work. Instead, we focus on proposing maps D and L representative of an *average listener* or *worst-case listener*, and motivated by the concept of *auditory event*, which is only related to the perceptual processing of the ear signals. The auditory scene is then regarded as the integration of different and separable auditory events.

The process of extracting auditory events from ear signals has been studied in the field of *auditory scene analysis* (ASA) [41], [39]. Although the psycho-acoustic and cognitive mechanisms involved are the focus of current research [1, Chapter 2.1], the quantitative modeling of the process has been carried out by the field of *computational auditory scene analysis* (CASA) [42]. The analysis of auditory scenes is a combination of bottom-up, or *signal driven*, and top-down, or *hypothesis driven*, processes [43]. In the former, the physical properties of the input signal, which are processed at the peripheral auditory system, serve as a basis for the formation of auditory events and are summarized under *primitive grouping cues* [41]. In the top-down process higher cognitive processes such as prior knowledge turn into play to determine which signal components the listeners attend to and how these components are assembled and recognized, involving processes that are referred to as *schema-based cues* [41]. Although a complete model of auditory events formation should also consider top-down processes, for simplicity we focus only on primitive grouping cues. These allow us to compare $\bar{u}_{(x,\theta)}$ and $\bar{u}_{0,(x,\theta)}$ by accounting mostly for the immediate psycho-acoustic peripheral processing.

The direct comparison between the auditory scene generated by $\bar{u}$ and $\bar{u}_0$ leads to the *full reference* (FR) *model*. In contrast, in the *internal references* (IR) *model* the auditory scene is compared to abstract representations in the listener's memory [44, Chapter 2.6]. Although FR models have been shown to have limitations [45], e.g., it is not always clear which binaural input signal is related to what the listener really desires to hear, we focus exclusively on them for simplicity. Otherwise schema based cues would be needed.

Spatial sound applications identify the most important *features* [2, Chapter 2] of the auditory scene in a multidimensional approach [44, Chapter 3.2]. In Letowski's simple model [46], the auditory scene is described in terms of *"loudness, pitch, (apparent) duration, spatial character (spaciousness), and timbre."* The last two are selected as the more important for spatial sound applications. The analysis of the most important features for spatial sound applications has been recently refined in [47], once again calling attention to timbral and spatial features. More specifically, azimuth localization and coloration are widely used features for the perceptual assessment of multi-channel reproduction systems [6], [10], [26]. Motivated by these studies, we focus on models that account only for coloration and azimuth localization. The trade-off is that these models can, at best, account for *plausibility* of the SAI, more than *authenticity*. Since the detailed biophysics of the phenomena are not necessary to model an accurate input-output relation, we only consider functional (phenomenological) models instead of physiological (biophysical) ones [48, Chapter 1.2].

A. Binaural azimuth localization

Azimuth localization is the estimation of the direction of arrival of the incoming sound in the horizontal plane. In concordance with Lord Rayleigh's *duplex theory* [49], the literature shows that interaural time differences (ITDs) are the primary azimuth localization cue at low frequencies [50] whereas interaural level differences (ILDs) become relevant at high frequencies as to resolve ambiguities in the decoding of ITDs [51] which appear over 1.4 kHz [52]. Consequently,

binaural models, i.e., models that use both the right and the left ear signals as inputs, are crucial, and most of these models for azimuth localization focus on the extraction of the ITDs. The cross-correlation between the left and right ear signals [53] is usually used to model the mechanism for extraction of the ITDs, e.g. [54], [55] and azimuth localization. Other models [51], [56] have appeared after the cross-correlation model was challenged by physiological findings [57]. The extraction of the ITDs using cross-correlation (or as in [51], [56]) and the extraction of ILDs, lead to dissimilarity maps D that do not satisfy (5). As we cannot model this effect within our framework at the moment, we look for alternative approaches in Section IV-C.

B. Binaural coloration

Coloration is commonly defined as timbre distortion [26], [1, Chapter 8.1] where timbre is the property that *"enables the listener to judge that two sounds which have, but do not have to have, the same spaciousness, loudness, pitch, and duration are dissimilar"* [46]. Although timbre could be quantified in a spectro-temporal space, the metric of the timbral space is not known and could be non-trivial [26].

Binaural perceptual effects such as binaural unmasking, spatial release from masking [58], and binaural decoloration [59], allow the auditory system to improve the quality of the perceived sound in terms of signal-to-noise ratio (SNR) identifiability and coloration. Even though these effects make defining an accurate binaural metric even more challenging, binaural detection and masking models have been developed in the literature [60]. They can be used to detect binaural timbral differences, and have been adapted to account for localization cues [61]. Furthermore, a model accounting for binaural perceptual attributes, such as coloration and localization, is proposed in [62]. More recently, a model for binaural coloration using multi-band loudness model weights to analyse the perceptual relevance of frequency components has been developed [63]. Similarly to the extraction of ITDs, these binaural methods once again lead to dissimilarity maps that do not satisfy (5). This leads us to consider monaural models in Section IV-C.

C. Monaural models

To our knowledge, the binaural models in the literature do not lead to dissimilarities satisfying (5). In contrast, under suitable assumptions, monaural models, i.e., models that need just one ear signal as input, do. Furthermore, they can be applied independently over each ear, to then use a *worst-case scenario* methodology [64] to extend them to binaural signals. We follow this approach and focus on monaural models as surrogates to capture coloration effects. Monaural spectral localization models have been developed for localization across the sagittal plane, and also for localization across the azimuthal plane [65], but, to our knowledge, they do not satisfy (5).

Models to detect monaural distortion aim to determine when two audio signals $s_0 = s_0(t)$ and $s = s(t)$ are perceived as different, and how this perception degrades as a function of the dissimilarities between s_0 and s. To achieve this, two main ideas are used for the estimation of audible distortions: the masked threshold and the comparison of internal representations [64].

The masked threshold compares the error signal $\varepsilon = s - s_0$ against s_0 using a perceptual distortion function $D^\star(\varepsilon, s_0)$. The error is assumed to be inaudible if this value is less than a fixed masking threshold [66], [67]. The comparison of internal representations leverages a model for an *internal representation* $s \mapsto I_R(s)$ resulting from the signal transformations performed in the ear. The internal representations are compared using an *internal detector* $(I_R(s), I_R(s_0)) \mapsto D^\star(I_R(s), I_R(s_0))$ and the difference between the signals is assumed to be perceptible if this value exceeds a given threshold [68], [64]. These studies do not provide analytical expressions that satisfy (5) for the representation nor for the internal detector. An approximation yielding such expressions is given in [69]; another simplified model is developed in [70].

The models developed in [66], [69], [70] yield monaural dissimilarity maps D that satisfy (5). These methods can be represented as

$$D^{\mathrm{m}}(s, s_0) = B_1(s - s_0) + \ldots + B_{n_b}(s - s_0) \tag{9}$$

where $B_1, \ldots B_{n_b}$ are filters of the form

$$B_k(s - s_0) = \int_{\mathbb{R}} \left| \int_{\mathbb{R}} K_{B_k}(t, t')(s - s_0)(t')\, dt' \right|^2 dt \tag{10}$$

for a suitable function K_{B_k} representing a time-variant or time-invariant filter that may depend on s_0 itself. In [66], [70] the filters B_k represent the auditory distortion over the k-th auditory filter of the cochlea, whereas in [69] the sum reduces to only one locally time invariant filter that accounts for the whole auditory distortion. Although monaural, these models can be used for binaural signals $\bar{s}$ by taking the worst distortion between ear signals [64]

$$D^{\mathrm{b}}(\bar{s}, \bar{s}_0) = \max\{D^{\mathrm{m}}(s^\ell, s_0^\ell), D^{\mathrm{m}}(s^r, s_0^r)\}. \tag{11}$$

D. Discomfort

To model the loudness discomfort L we consider empirical evaluations of discomfort. This is a simplification motivated by computational simplicity and also by a small number of comprehensive studies on the subject. Empirical thresholds for loud discomfort levels for sinusoidal signals over a finite set of frequencies have been defined in the literature, e.g. in [71], [72]. Naturally, for a sinusoidal signal of frequency f_k these can be expressed with the monaural expression

$$Q_k(s) = \int_{\mathbb{R}} |\widehat{s}(f)|^2 \, \rho_k(f) df, \tag{12}$$

where $\rho_k(f) = (\gamma(f)/\eta_k)^2$, γ is a narrow band-pass filter centered around f_k and $\eta_k \in \mathbb{R}_+$ is the discomfort threshold at f_k. Finally, for binaural signals, these models can be applied taking the worst discomfort between ears as in (11).

V. Implementation

We implement SWEET-ReLU to approximate a sound wave generated by a monopole emitting close to a single frequency f_0. We call this a (pseudo) sinusoidal source. We consider this to be a proof-of-concept implementation to illustrate the performance of the method.

A. Implementation of acoustic framework

The original sound wave u_0 is assumed to be emitted by a monopole at $x_0 \in \mathbb{R}^3$. Each loudspeaker is assumed to radiate as a monopole. Hence, the sound wave u generated by the array is given by (1) with $\widehat{\alpha}_k(f) = a_k\, e^{-(f-f_0)^2/2\sigma^2}$ for coefficients $a_k \in \mathbb{C}$ and a fixed spectral localization parameter $\sigma \ll 1$.

B. Implementation of the perceptual framework

Monaural distortion detectability methods cannot represent the necessary features to correctly define an auditory illusion map as described in Section IV. Hence, they may not be optimal when modelling binaural perception. For instance, they cannot represent explicitly any type of azimuth localization, nor binaural coloration effects. However, they can represent monaural coloration effects, and some yield dissimilarity maps satisfying (5). For this reason, we use a monaural model as a proof-of-concept. We use van de Par's spectral psycho-acoustic model [66]. Although it is suboptimal when modeling temporal masking effects, the signals we consider are stationary, whence temporal masking is almost non-existent.

This monaural model can be applied to binaural signals by using the worst-case as in (11). For $x \in \Omega$ and $\theta \in \Theta_x$ we apply this model to the left and right ear signals $\bar{u}(x,\theta)$ and $\bar{u}_0(x,\theta)$. As van de Par's model can be represented by time-invariant filters in (10), for $s \in \{\ell, r\}$ we have that

$$B_j u^s(x,\theta) = \int_{\mathbb{R}} |(\widehat{u}^s - \widehat{u}_0^s)(f,x,\theta)|^2 \rho_{B_j}(f,x,\theta)\, df$$

where ρ_{B_j} depends on u_0 as

$$\rho_{B_j}(f,x,\theta) = \frac{w_j(f)}{C_A + \int_{\mathbb{R}} |\widehat{u}_0^s(f,x,\theta)|^2 w_j(f) df}.$$

The constant $C_A > 0$ limits the perception of very weak signals in silence. The weight w_j is defined as $w_j := |\eta\gamma_j|^2$ where

$$\log_{10}\eta(f) = C_{\eta,0} - C_{\eta,1} f^{-0.8} - C_{\eta,2}(f - 3.3\times 10^3)^2 + C_{\eta,3} f^4$$

models the outer and middle ear as proposed by Terhardt [73] with $C_{\eta,0} = 4.69$, $C_{\eta,1} = 18.2 \times 10^{1.4}$, $C_{\eta,2} = 32.5 \times 10^{-7}$ and $C_{\eta,3} = 5 \times 10^{-16}$, and

$$\gamma_j(f) = \left(1 + \left(\frac{945\pi(f - f_j)}{48\mathrm{ERB}(f_j)}\right)^2\right)^{-2}$$

models the filtering property of the basilar membrane in the inner ear at the center frequency f_j, where the Equivalent Rectangular Bandwidth (ERB) of the auditory filter centered at f_j is $\mathrm{ERB}(f_j) = 24.7(1 + 4.37 \times 10^{-3} f_j)^{-1}$ as suggested by Glasberg and Moore [74]. The center frequencies f_j are uniformly spaced on the ERB-rate scale $\mathrm{ERBS}(f) = 21.4 \log(1 + 4.37 \times 10^{-3} f)$. In van de Par's model, the distortion is noticeable when its metric is greater or equal to 1. Hence, the monaural dissimilarity map becomes

$$\begin{aligned} D^{\mathrm{m}}_{u^s,u_0^s}(x,\theta) &= -1 + C_0 \sum_{j=1}^{n_b} \frac{\int_{\mathbb{R}} |(\widehat{u}^s - \widehat{u}_0^s)(f,x,\theta)|^2 w_j(f) df}{C_A + \int_{\mathbb{R}} |\widehat{u}_0^s(f,x,\theta)|^2 w_j(f) df} \\ &\approx -1 + C_0' \sum_{j=1}^{n_b} \frac{|(\widehat{u}^s - \widehat{u}_0^s)(f_0,x,\theta)|^2 w_j(f_0)}{C_A + w_j(f_0)|\widehat{u}_0^s(f_0,x,\theta)|^2} \end{aligned}$$

where $C_0' = 2^{1/4}\pi^{1/2}\sigma C_0$ and we used the approximation for (pseudo) sinusoidal signals

$$\int_{\mathbb{R}} \varphi(f) |\widehat{u}_0^s(f,x,\theta)|^2\, df \approx \sqrt{2^{1/2}\pi}\sigma\varphi(f_0)|\widehat{u}_0^s(f_0,x,\theta)|^2$$

when $\sigma \ll 1$. The constants C_0' and C_A are defined as suggested in [66]. This accounts for the absolute threshold of hearing and the just-noticeable difference in level for sinusoidal signals, which gives, $C_0' \approx 1.555$ and $C_A \approx 4.481$ when considering $n_b = 100$ as the number of center frequencies, and $f_1 = 20, f_{n_b} = 10^3$ as the first and last center frequency. The worst-case extension of this perceptual dissimilarity to binaural signals is given by

$$D_{\bar{u},\bar{u}_0}(x,\theta) = \max\{D^{\mathrm{m}}_{u^\ell,u_0^\ell}(x,\theta), D^{\mathrm{m}}_{u^r,u_0^r}(x,\theta)\}.$$

To model the loudness discomfort we use the experimental results in [71] about the discomfort caused by sinusoidal signals. We interpolate the data in this study with cubic splines with natural boundary [75, Section 8.6] to obtain a function $\eta_L > 0$. Therefore, following (12) we consider

$$\begin{aligned} L^{\mathrm{m}}_{u^s}(x,\theta) &= -1 + C_1 \int_{\mathbb{R}} |\widehat{u}^s(f,x,\theta)|^2 \rho_L(f) df \\ &\approx -1 + C_1' |\widehat{u}^s(f_0,x,\theta)/\eta_L(f_0)|^2 \end{aligned}$$

where $\rho_L = (\gamma_0/\eta_P)^2$ and the same approximation holds by the same arguments as before. Naturally, $C_1' = 1$, as the empirical thresholds in [71] are attained when $L^{\mathrm{m}} u^s(x) \le 0$. Then, the worst-case extension to binaural signals is

$$L_{\bar{u}}(x,\theta) = \max\{L^{\mathrm{m}}_{u^\ell}(x,\theta), L^{\mathrm{m}}_{u^r}(x,\theta)\}.$$

Although this implementation is intended to study single-frequency signals, it can be generalized to multifrequency signals. The generalization of D^{m} is direct and it comprises additional terms in the sum. The generalization of L^{m} to a multifrequency signal could sum the discomfort associated to each frequency, similarly to the way the auditory filter errors are summed in D^{m}, or it could use an integrating function as indicated in Section IV-D. These are simple heuristics for a proof-of-concept study, and their effectiveness should be validated experimentally.

C. Discretization

In a typical experiment, listeners are seated in a room, and their locations and orientations within the room are determined in advance. In this case, a simplification consists in defining the sweet spot in terms of the number listeners for which the SAI is achieved. If the listeners can be located at finite number

of points $z_1, \dots, z_{n_\ell} \in \Omega$ then we can model the weighted area of the sweet spot as

$$\mu(\mathcal{S}(\bar{u})) = \sum_{\ell=1}^{n_\ell} T_D \bar{u}(z_\ell), \tag{13}$$

which is equivalent to assuming μ is an atomic measure. In this case, every component of the proof-of-concept implementation can be evaluated either in closed-form, as is the case of the weighted area or the Green function of the loudspeakers, or can be approximated to very high-accuracy, as is the case of the head-related transfer functions (see Section VI for details).

The approximation (13) can be used when there is a discrete number of locations for listeners in a room. In other applications where it is necessary to control a continuum, e.g., when listeners may move across the room, quadrature rules must be used. In this case, the approach to solve (P_0) follows an approximate-then-discretize approach, and numerical errors may have an effect on the sweet spot computed in practice.

VI. Experiments

For the experiments we compare the performance of our method with the state-of-the-art methods WFS, NFC-HOA and L^2-PMM in terms of its azimuth localization and coloration performance, as they are the main features of the auditory scene for spatial sound. See [76] for the implementation in Python. We use Dietz's model to measure binaural azimuth localization [51] and McKenzie's model for binaural coloration [63]. The setup for the numerical experiments consists of an equispaced arrangement of 20 loudspeakers lying on a circle of radius 2.5 m and at $\pi/4 \approx 0.785$ m from each other. The region of interest Ω is a concentric circle of radius 2.4975 m (Figs. 1j, 2j). The speed of sound is $c = 343$ m/s. Two instances of this setup were evaluated: the *near-field instance*, where the source is outside the arrangement at 5 m of its center with $a_0 = 68$ dB, and the *focus-source instance*, where the source is inside the arrangement at 0.82 m of its center with $a_0 = 60$ dB (Figs. 1a, 2a). In both cases, $f_0 = 343$ Hz. To construct the perceptual maps D and L we assume that a listener located at x looks at the virtual sound source, that is, $\Theta_x = \{\text{ang}(x_0 - x)\}$.

The SWEET-ReLU algorithm and the L^2-PMM method were implemented in Python 3.8 using the CVXPY package, version 1.1.15 [77], [78] and MOSEK, version 9.3.6 [79]. The simulations of 2.5D NFC-HOA and 2.5D WFS were done with the Sound Field Synthesis Toolbox (SFST), version 3.2 [80], except for the focus source 2.5D NFC-HOA simulations, which were done following the *angular weighting approach* [81] as proposed in [82, Chapter 5.6.2]. The HRTFs used to simulate $\bar{u}$ and $\bar{u}_0$ were constructed as the circulant Fourier transform of the elements of the TU-Berlin HRIR free data base [83]. Dietz's and McKenzie's models were implemented using Matlab 2022a with the Auditory Modelling Toolbox (AMT), version 1.1 [84].

For the implementation of the HRTFs, the 3 meters radial distance HRIRs of the data set were radially extrapolated using delay and attenuation, according to the map $d \mapsto (3/d)\text{HRIR}(t - (d-3)/c)$, where d is the desired radial distance. For short distances, e.g. less than 1 m, ILDs vary significantly with distance [83]. Hence, the experiments might be enhanced by using a complete HRTF data set.

For the implementation of Dietz's model, since we treated (pseudo) sinusoidal signals, the interaural phase differences (IPDs) of the reproduced signals were extracted manually as $\text{IPD}(\bar{u}(x,\theta)) = \arg(u^\ell(x,\theta)u^r(x,\theta)^*)$. Since $f_0 = 343$ Hz, the IPD to estimated azimuth localization is an injective map [51]. Thus, from the relation $\text{ITD} = \text{IPD}/(-2\pi f_0)$ [51], the estimated azimuth localization was obtained by plugging the ITDs into the `itd2angle.mat` AMT function, which uses the `itd2angle_lookuptable.mat` AMT table. The latter table is based on Dietz's model and is constructed with the same HRTFs that we consider for the simulation of $\bar{u}$ and $\bar{u}_0$. For the implementation of McKenzie's model, the binaural signals were transformed to time-domain using a sampling frequency of 44100 Hz and a number of samples of 256. For the implementation of the SWEET method, we have chosen ε_i adaptively with percentile $p = 99$. For SWEET and L^2-PMM a uniform discretization of 2348 points was used for Ω at a distance of at most 0.09 m, achieving more than 30 points per wavelength.

	SWEET	NFC-HOA	WFS	L^2-PMM
NF CSS	70.2%	60 %	55%	3.3%
NF LSS	68.4%	42.9%	47.9%	52.3%
FS CSS	58.2%	29.3%	0%	5%
FS LSS	54 %	42.9%	16.2%	40.1%
FS LSS (H)	50.8%	37.8%	12.5%	28.9%

TABLE I: Localization (LSS) and coloration (CSS) sweet spots over Ω fractions in near-field (NF) and focus-source (FS) instances. (H) disregards the convergent halfspace.

To compare the performance of the methods, we measure the size of the *localization sweet spot* (LSS) and *coloration sweet spot* (CSS). The former is the region where the perceived azimuth localization measured by Dietz's model deviates no more than 5 degrees from the desired one, whereas the latter is the region where the coloration measured by McKenzie's model is lower or equal than 13 sones. As discussed in [63], a coloration lower than 13 sones is strongly correlated with empirical MUSHRA tests with more than 80 out of 100 points. It should be noted that, since we analyse sinusoidal signals of frequency 343 Hz, in these experiments the CSS and LSS are constituted by the points where the interaural amplitude or phase (respectively) of the binaural signal are correctly reconstructed. For this reason, we also show the IPDs over Ω.

The LSS, CSS and IPDs generated by each method for the near-field and focus-source instances are shown in Fig. 1 and 2 respectively. The size of the LSS and CSS is shown in Table I. The blue zones of Figs. 1k-r, Figs. 2k-r represent the sweet spots of each case. The estimated localization in Figs. 1k-n, Figs 2k-n is shown for deviations of 0 to 90 degrees from the desired one. Greater deviations are represented by a dot with no direction. Since the potential listeners are looking to the virtual source, the desired IPDs are equal to 0 in each spatial point by symmetry. Furthermore, as in Dietz's model the perceived azimuth localization is a function of the IPDs

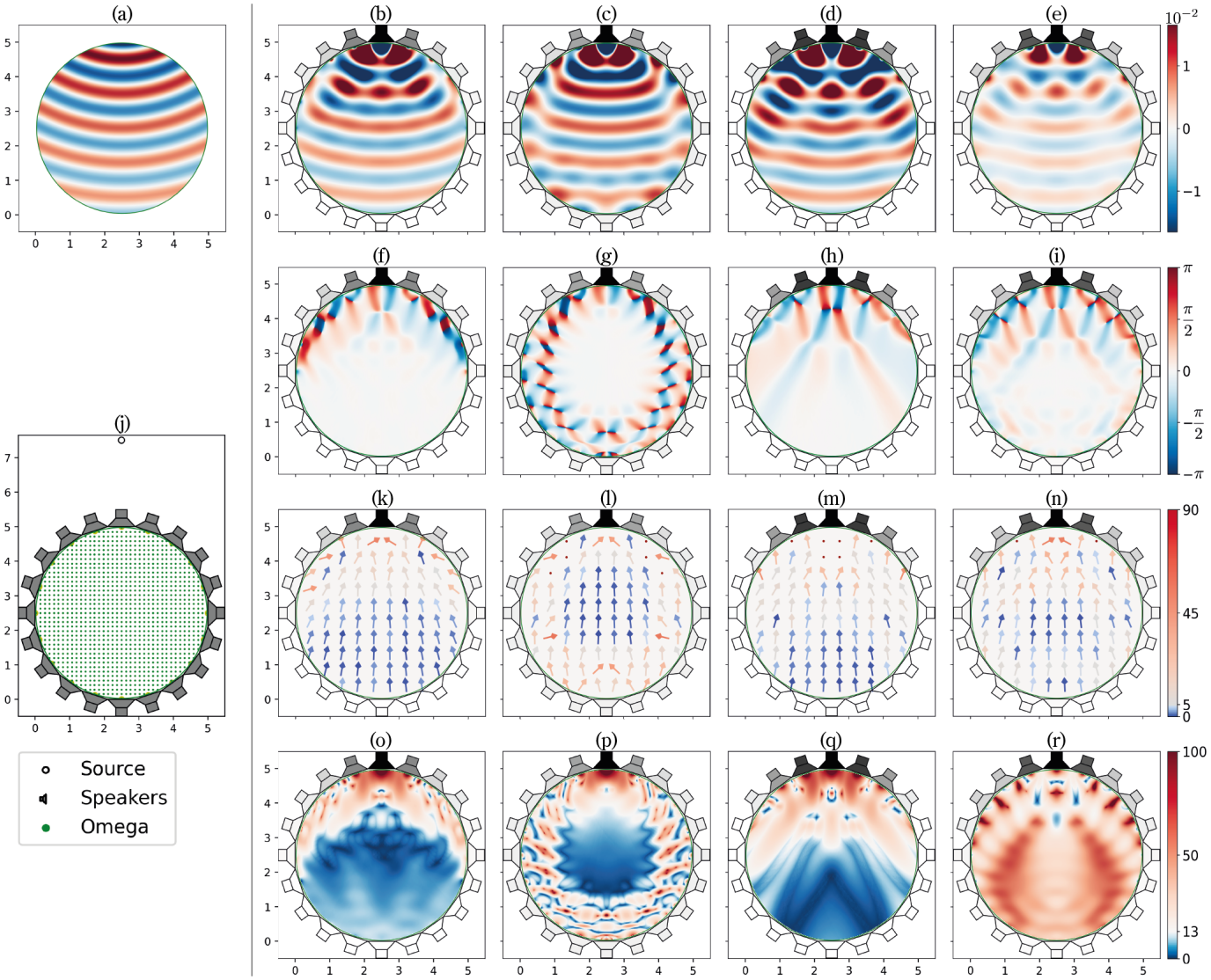

Fig. 1: Near-field instance. *Left panel:* $\hat{u}_0(f_0)$ (real part), (j) Instance configuration (spatial units in meters). *Right panel rows:* Near-field $\widehat{u}(f_0)$ (real part); IPD($\bar{u}$); Dietz's azimuth localization, where the direction of the arrows is the perceived localization whereas the color is the deviation in degrees of the perceived localization from the desired one; McKenzie's coloration (sones). *Right panel columns:* SWEET-ReLU, NFC-HOA, WFS, L^2-PMM. The shading of the speakers is proportional to their gain.

at low frequencies, the regions where the IPD deviates from 0 at Figs. 1f-i and Figs. 2f-i strongly correlates with the LSS in Figs. 1k-n and Figs 2k-n, respectively.

For the near-field instance, both the LSS and CSS generated by our method are more than 20 and 10 points (respectively) larger than that generated by any other method. The LSS and CSS generated by NFC-HOA (Figs. 1l, 1p) are centered, whereas that generated by WFS (Figs. 1m, 1q) are localized farther away from the source. This is correlated with their degradation of the sound field, which is consistent with the analysis in [14]. Moreover, their LSS are consistent with the empirical results exposed in [25]: the perceived localization for NFC-HOA degrades away from the center, whereas for WFS the perceived localization is fairly good over almost all the listening region. However, we believe that the perceived localization for WFS and NFC-HOA behaves slightly worse here than in [25] because we analyze sinusoidal signals instead of Gaussian white noise, which has uniform spectral content. In contrast, the LSS and CSS generated by our method (Figs. 1k, 1o) behave like those generated by WFS, but almost encompasses those generated by NFC-HOA. The LSS of L^2-PMM (Fig. 1n) is larger than that of WFS and NFC-HOA, but its CSS (Fig. 1r) is almost negligible. This is consistent with the sound wave u produced with L^2-PMM (Fig. 1e) as the phase of the signals is fairly well reconstructed (Fig. 1i), whereas its amplitude is too small.

In the focus-source instance, due to reasons of causality, theoretically any method can achieve a correct reproduction of the direction of propagation of u_0 only in one half-space defined by $\{x_i\}_{i=0}^{n_s}$, where the sound field diverges from the focus-source position [85]. In the other half-space the reproduced wave field converges towards the location of the virtual source. As shown in Figs. 2k-n, the LSS of all the methods comprise a portion of the converging part of the

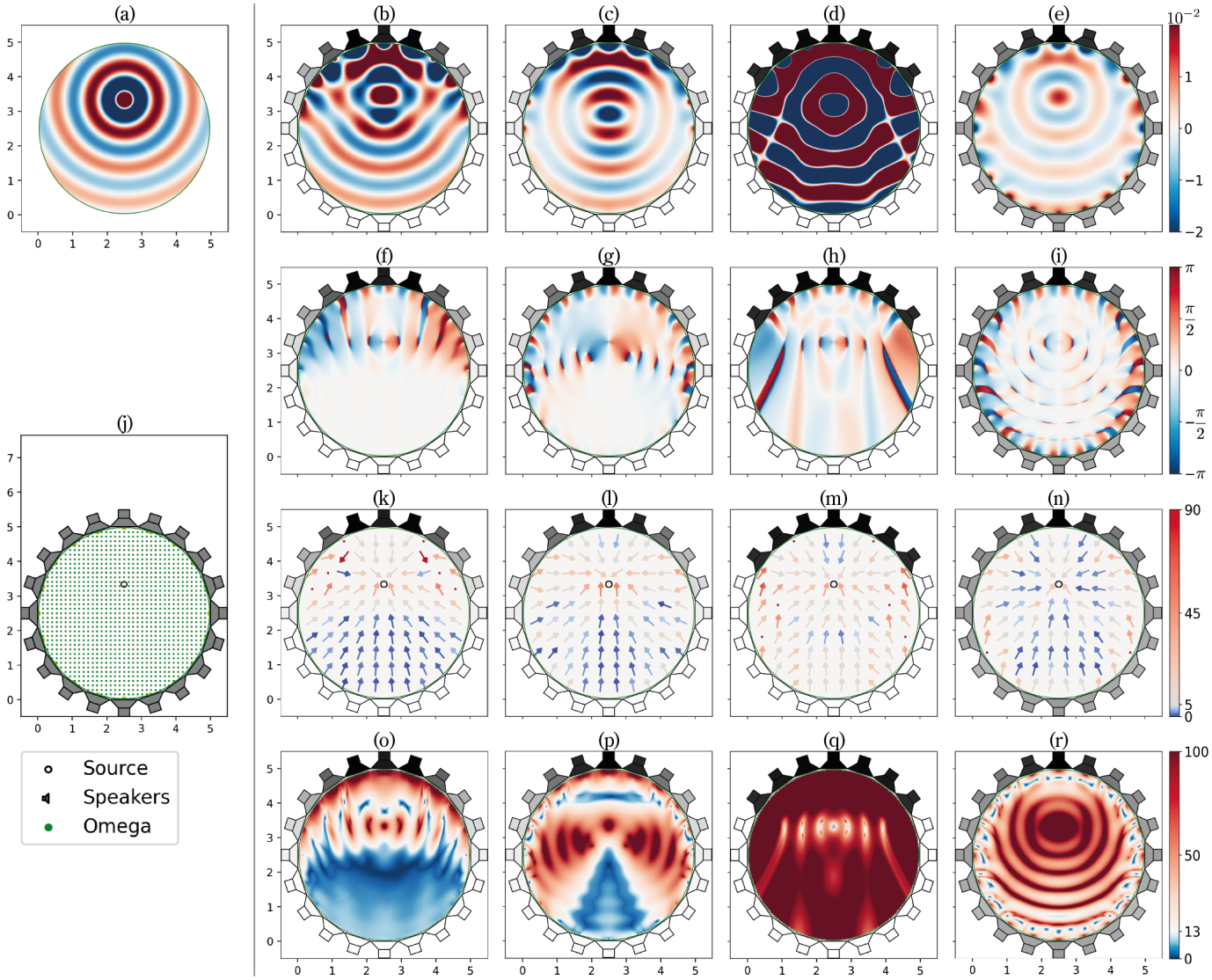

Fig. 2: Focus-source instance. *Left panel:* (a) $\widehat{u}_0(f_0)$ (real part), (j) Instance configuration (spatial units in meters). *Right panel rows:* Near-field $\widehat{u}(f_0)$ (real part); IPD($\bar{u}$); Dietz's azimuth localization, where the direction of the arrows is the perceived localization whereas the color is the deviation in degrees of the perceived localization from the desired one; McKenzie's coloration (sones). *Right panel columns:* SWEET-ReLU, NFC-HOA, WFS, L^2-PMM. The shading of the speakers is proportional to their gain.

sound field. This is possible because the interaural phase of the binaural sinusoidal signals is correctly recreated at those points, as shown in Figs. 2f-i. However, we believe that in more complex scenarios, involving multi-frequency signals and allowing the listener to turn her head, e.g., considering a larger Θ, it would not be possible to recreate correctly and consistently the localization illusion. As shown in [82, Chapter 5.6] *"for listeners located in the converging part of the sound field, the perception is unpredictable since the interaural cues are either contradictory or change in a contradictory way when the listener moves the head."* Table I contains the fraction of the LSS over Ω for the focus-source instance considering all the points of the converging half-space as incorrectly reconstructed.

For the focus-source instance, the LSS and CSS generated by our method (Figs. 2k, 2o) is approximately 10 and 30 (respectively) points larger than those generated by other methods. The LSS and CSS generated by NFC-HOA (Figs. 2l, 2p) is concentrated in a limited region at the diverging part and around a vertical line that passes through x_0. The LSS generated by WFS (Fig. 2m) is almost contained in the same vertical line, although the error in the localization reconstruction is below 15 degrees in a larger area. The CSS generated by WFS (Fig. 2q) is almost empty as the resulting u has a large amplitude. This suggests that a focus-source formulation for WFS needs a factor for amplitude normalization. The LSS of L^2-PMM (Fig. 2n) has almost the same size as that of NFC-HOA, but its CSS (Fig. 2r) is almost negligible. This is consistent with the sound wave u produced with L^2-PMM (Fig. 2e) as the phase of the signals is fairly well reconstructed (Fig. 2i), whereas its amplitude is too small.

The LSS and CSS generated by our method (Figs. 2k, 2o) comprises almost all the divergent part. This highlights one of the advantages of the greedy approach of SWEET-ReLU: it is capable to detect the direction of u_0 over Ω in its first iterations, to then prioritize the part of Ω where a good fit to $\bar{u}_0$ can be obtained. This is a possible explanation for the amplitude mismatch of L^2-PMM both in the near-field (Fig. 1e) and the focus-source (Fig. 2e) instances: just minimizing the square of the spatial errors strongly penalizes the spatial points where the amplitude is very large and difficult to reconstruct, i.e., near the loudspeakers. Then, L^2-PMM finds a solution where the amplitude error at those points is not too large, leading to a small overall amplitude. This suggests that the application of spatial weighting matrix could enhance the amplitude matching of L^2-PMM.

VII. Discussion

Our results show the SWEET-ReLU method yields state-of-the-art results in standard numerical experiments with our proof-of-concept implementation. We believe the performance in these experiments is representative of what we would observe when using more complex pyscho-acoustic models for the perceptual dissimilarity and the loudness discomfort. A key component of our method is the perceptual dissimilarity D. Although its form in our proof-of-concept implementation is quite flexible, it does not account for spatialization and other binaural effects. Finding a model to account for these effects such that D satisfies (5) is the subject of future research. It should be noted that, as it is shown in [86], the overall quality of a spatial sound system can be explained to 70% by coloration or timbral fidelity, which can be partially characterized by monaural effects, and 30% by spatial fidelity, which needs to be characterized by binaural effects.

Furthermore, our proof-of-concept experiments show that even though the perceptual model we use is an extension of a monaural model using a worst-case approach, it still is able to perform better than state-of-the-art methods in terms of localization and coloration. This suggests our implementation with this model is able to capture correctly some of the spatial properties of the auditory scene, even though these properties are not explicitly in the model. This might be explained because the coloration and localization is strongly dependant of the amplitude and phase of listener binaural signals, which is controlled by our implementation of a binaural extension of an amplitude and phase-sensitive monaural model.

Although our implementation assumes the loudspeakers and the sources are monopoles, we believe our method can be readily implemented in real settings with non-trivial sound sources. For instance, reverberation, different radiation patterns for the loudspeakers, and other time-invariant effects can be incorporated by modifying the Green function G_k in (2) and the transfer functions in (4) accordingly.

Even though we have not fully developed a theory for the convergence of SWEET-ReLU, our numerical experiments show that the method converges to reasonable results in practice. Furthermore, our proof-of-concept implementation avoids any potential issues arising from the discretization of the models, either due to numerical computation of the Green function or transfer functions, or to the discretization of the integral that defines the weighted area. Further analysis about this point will be the subject of future work.

Finally, our method addresses the two fundamental drawbacks pointed out in [82, Chapter 1.4] about the sound field synthesis numerical methods. First, the optimization criteria of our method is based on perceptual features. Second, our model is aware of fundamental physical restrictions of the secondary source setup under consideration such as 2.5-dimensionality and the spatial discrete property of real-world setups.

VIII. Conclusion

In this work, we introduced a theoretical framework for spatial audio perception that allows the definition of a perceptual sweet spot, that is, the region where the spatial auditory illusion is achieved when approximating one sound wave by another. Furthermore, we developed a method that finds an approximating sound wave that maximizes this sweet spot while guaranteeing no loudness discomfort over a spatial region of interest. We provided a theoretical analysis of the method, and an efficient algorithm, the SWEET-ReLU algorithm, for its numerical implementation. In a proof-of-concept implementation using monopoles emitting (pseudo) sinusoidal signals, our method successfully captures some of the spatial properties of the auditory scene, such as localization and coloration, even though these properties are not explicitly in the model. We believe our method is a first step towards a novel approach for spatial sound with loudspeakers, bridging the gap between methods based on perceptual principles, and sound field synthesis methods.

Appendix A
The SWEET method

A. Preliminaries

We let $L^2(\mathbb{R})$ be the space of (equivalence classes of) *complex-valued functions* that are modulus-square integrable with respect to the Lebesgue measure on $\mathbb{R}$ and for a set S we let $C^0(S)$ be the set of complex-valued continuous functions defined on S. Let

$$X_{\text{ES}} := \{(u^\ell, u^r) : u^s \in L^2(\mathbb{R}),\, s \in \{\ell, r\}\}$$

be the space of *pairs of signals* or *ear signals*. When endowed with

$$\|\bar{u}\|^2_{X_{\text{ES}}} = \int_{\mathbb{R}} |u^\ell(t)|^2\, dt + \int_{\mathbb{R}} |u^r(t)|^2\, dt$$

it becomes a complete metric space. Define the set $Z := \cup_{x\in\Omega}\{x\} \times \Theta_x$ and endow it with the subspace topology in $\Omega \times \mathbb{R}^3$ (or $\Omega \times \mathbb{R}^2$). Define the space

$$W := \{\bar{u} : Z \to X_{\text{ES}} : \bar{u} \text{ continuous and bounded}\}$$

of *spatial distributions of pairs of signals*; these are not equivalence classes. If $\bar{u} \in W$ then $\bar{u}_{(x,\theta)} \in X_{\text{ES}}$ for every $(x, \theta) \in Z$. When endowed with the norm

$$\|\bar{u}\|_W := \sup_{(x,\theta)\in Z} \|\bar{u}_{(x,\theta)}\|_{X_{\text{ES}}}$$

the space W is complete. The Fourier transform is an isometry in $L^2(\mathbb{R})$. We define $\mathcal{F} : W \to W$ as $\mathcal{F}\bar{u}_{(x,\theta)} = (\widehat{u}^{\ell}_{(x,\theta)}, \widehat{u}^{r}_{(x,\theta)})$. It can be verified $\mathcal{F}$ is an isometry in W. Let $I_S \subset \mathbb{R}$ and let $\gamma_{\max} > 0$. The set of *admissible audio signals* driving each loudspeaker is

$$X_{\text{AS}} := \{\alpha \in L^2(\mathbb{R};\mathbb{C}) : \widehat{\alpha}|_{I_S^c} = 0, \|\widehat{\alpha}\|_{L^2} \leq \gamma_{\max}\}$$

where $|$ denotes restriction; they are bandlimited to I_S and have norm bounded by $\gamma_{\max}$. The sound waves in (4) belong to

$$W_S := \left\{ \mathcal{F}^{-1}\left(\sum_{k=1}^{n_s} \widehat{\alpha}_k H_k^r, \sum_{k=1}^{n_s} \widehat{\alpha}_k H_k^{\ell} \right) : \alpha_k \in X_{\text{AS}} \right\}.$$

To quantify the area of the sweet spot in (7) we use a finite Borel measure μ on Ω [87, Section 1.2]; we usually suppose μ is absolutely continuous with respect to the Lebesgue measure or atomic. If $\bar{u}$ is the pair of signals generated by the array, then $\mu(\mathcal{S}(\bar{u}))$ is the weighted area of the sweet spot $\mathcal{S}(\bar{u})$. We consider the space $L^\infty_\mu(\Omega)$ of (equivalence classes of) real-valued Borel measurable functions that are bounded μ-a.e. [87, Section 3.3]. Frequently used operations, such as the sum, supremum or infimum of functions, the integral, and inequalities, are well-defined for such equivalence classes of measurable functions. If $v \in L^\infty_\mu(\Omega)$ then $\mu(\{x \in \Omega :\ v(x) \geq 0\})$ is independent of the representative used.

Assumption 1. (i) Ω is compact. (ii) $\Theta_x \neq \emptyset$ for every $x \in \Omega$. (iii) Z is compact. (iv) I_S is compact. (v) The functions H_k^{ℓ}, H_k^r are continuous and bounded on $I_S \times Z$. (vi) The dissimilarity map $D : W \times W \to C^0(Z)$ is continuous and convex on its first argument. (vii) The discomfort map $L : W \to C^0(Z)$ is continuous and convex. (viii) $\bar{u}_0 \in W$.

The model proposed in Section II-B is well-defined.

Proposition 1. *Under Assumption 1 the following assertions are true: (i) The set W_S is convex and compact in W. (ii) The map $T_D : W_S \to L^\infty_\mu(\Omega)$ is continuous, and for $x \in \Omega$ the map $\bar{u} \to T_D\bar{u}(x)$ in* (6) *is convex. (iii) The map $\mu \circ \mathcal{S} : W \to \mathbb{R}$ is well-defined, that is, its values do not depend on the choice of representative of $T_D\bar{u}$. (iv) The set $\mathcal{P}$ in* (8) *is convex and closed in W.*

We defer the proof to Appendix A-H. Although the feasible set for (P_0) is compact, the objective depends on the properties of the *set-valued function* $u \rightrightarrows \mathcal{S}(u)$. As studying these properties and minimizing $\mu \circ \mathcal{S}$ is potentially challenging, we propose an approximation to (P_0) that can be analyzed and solved with standard methods.

B. The layer-cake representation

We approximate the area of $\mathcal{S}(u)$ using the *layer-cake representation*.

Assumption 2. $\varphi : \mathbb{R} \to \mathbb{R}$ is absolutely continuous, bounded, non-negative and such that $\varphi(t) = 0$ for $t < 0$ and $\|\varphi\|_{L^1} = 1$.

For $\varepsilon > 0$ let φ_ε denote $\varphi_\varepsilon(t) = \varphi(t/\varepsilon)/\varepsilon$ and define

$$\Phi_\varepsilon(t) = \int_{-\infty}^{t} \varphi_\varepsilon(s)\, ds.$$

Since Φ_ε is continuous, the composition $\Phi_\varepsilon \circ v$ is well-defined as an element in $L^\infty_\mu(\Omega)$ for any $v \in L^\infty_\mu(\Omega)$. Similarly,

$$A_\varepsilon(v) := \int_\Omega \Phi_\varepsilon(v(x))\, d\mu(x).$$

is well-defined for any $v \in L^\infty_\mu(\Omega)$. By Assumption 2 Φ_ε is non-decreasing, whence $v_1 \leq v_2$ implies $A_\varepsilon(v_1) \leq A_\varepsilon(v_2)$, i.e., A_ε is non-decreasing.

Proposition 2. *Under Assumption 2, for every $v \in L^\infty_\mu(\Omega)$ we have*

$$\lim_{\varepsilon \downarrow 0} A_\varepsilon(v) = \mu(\{x \in \Omega : v'(x) > 0\}).$$

Proof of Proposition 2. Let $t \geq 0$, let v' be a representative, and let $\{\varepsilon_n\}_{n\in\mathbb{N}}$ be non-negative and monotone decreasing to zero. Define $V_{t,n} := \{x \in \Omega : v'(x) \geq \varepsilon_n t\}$ and note that $V_{t,n} \subseteq V_{t,n+1}$. Define $V := \bigcup_{n>0} V_{t,n} = \{x \in \Omega : v'(x) > 0\}$ and $h_n(t) = \varphi(t)\mu(V_{t,n})$. Note h_n is measurable for every n as $t \mapsto \mu(V_{t,n})$ is monotone. Then $h_n(t) \uparrow \varphi(t)\mu(V)$ as $n \to \infty$ by continuity from below [87, Proposition 1.2.5]. Since Ω is bounded, $v \in L^1_\mu(\Omega)$ and v' is absolutely integrable. By Fubini's theorem,

$$\begin{aligned}
A_{\varepsilon_n}(v) &= \int_\Omega \int_{-\infty}^{v'(x)} \varphi_{\varepsilon_n}(t) dt d\mu(x) \\
&= \int_{\mathbb{R}} \varphi_{\varepsilon_n}(t) \int_\Omega \chi_{\{x\in\Omega : v'(x) \geq t\}}(t,x)\ d\mu(x) dt \\
&= \int_{\mathbb{R}} \varphi_{\varepsilon_n}(t) \mu(\{x \in \Omega : v'(x) \geq t\}) dt \\
&= \int_0^\infty \varphi(s)\mu(V_{s,n}) ds \xrightarrow{n\to\infty} \mu(\{x \in \Omega : v'(x) > 0\})
\end{aligned}$$

where we used the change of variables $s = t/\varepsilon_n$ and the monotone convergence theorem [87, Theorem 2.4.1]. As $\{\varepsilon_n\}_{n\in\mathbb{N}}$ is arbitrary, the claim follows. □

For $\bar{u} \in W_S$ and $\varepsilon > 0$ sufficiently small we have by Propositions 1 and 2 that

$$\begin{aligned}
A_\varepsilon(T_D\bar{u}) &= \int_\Omega \Phi_\varepsilon(T_D\bar{u}(x))\, d\mu(x) \\
&\approx \mu(\{x \in \Omega : T_D\bar{u}(x) > 0\}) = \mu(\Omega) - \mu(\mathcal{S}(\bar{u})).
\end{aligned}$$

Consequently, $\mu(\mathcal{S}(\bar{u})) \approx \mu(\Omega) - A_\varepsilon(T_D\bar{u})$ and for every fixed $\bar{u} \in W_S$ we can approximate $\mu(\mathcal{S}(\bar{u}))$ by $A_\varepsilon(T_D\bar{u})$.

C. The variational problem

We propose to solve the surrogate problem

$$(P_\varepsilon) \quad \underset{\bar{u}\in W_S \cap \mathcal{P}}{\text{minimize}} \quad A_\varepsilon(T_D\bar{u}) \tag{14}$$

Proposition 3. *Suppose Assumptions 1 and 2 hold. The function $A_\varepsilon : L^\infty_\mu(\Omega) \to \mathbb{R}$ is continuous, and (P_ε) has at least one minimizer.*

Proof of Proposition 3. Let $\delta > 0$, let $v_0, v \in L^\infty_\mu(\Omega)$ be such that $\|v - v_0\|_{L^\infty_\mu} < \delta/2$, and let v' and v'_0 be representatives. There exists $\Omega^* \subset \Omega$ with $\mu(\Omega \setminus \Omega^*) = 0$ such that $|v'(x) - v'_0(x)| < \delta/2$ for $x \in \Omega^*$. Since φ_ε is non-negative

and bounded on the interval $[-\|v_0\|_{L^\infty_\mu} - \delta/2,\ \|v_0\|_{L^\infty_\mu} + \delta/2]$, for any $x \in \Omega^*$ we have that

$$|\Phi_\varepsilon(v'(x)) - \Phi_\varepsilon(v'_0(x))| \leq \int_{v'_0(x)-\delta/2}^{v'_0(x)+\delta/2} \varphi_\varepsilon(t)\, dt \leq c_{\varphi_\varepsilon}\delta$$

where $c_{\varphi_\varepsilon} > 0$ depends only on φ_ε. As the bound is independent of the choice of v', v'_0, $|A_\varepsilon(v) - A_\varepsilon(v_0)| \leq c_{\varphi_\varepsilon}\mu(\Omega)\delta$ whence A_ε is continuous. The existence of solutions follows from the compactness of $W_S \cap \mathcal{P}$. □

D. DC Formulation

To solve (P_ε) we first rewrite it equivalently as

$$(\tilde{P}_\varepsilon) \quad \begin{cases} \underset{\substack{\bar{u}\in W_S\cap\mathcal{P} \\ v\in L^\infty_\mu(\Omega)}}{\text{minimize}} & A_\varepsilon(v) \\ \text{subject to} & T_D\bar{u} \leq v. \end{cases} \tag{15}$$

We interpret the auxiliary variable v as an overestimate of the auditory illusion map over Ω.

Proposition 4. *Under Assumptions 1 and 2, if $\bar{u}^\star$ is an optimal solution to (P_ε) then $(\bar{u}^\star, T_D\bar{u}^\star)$ is an optimal solution to $(\tilde{P}_\varepsilon)$. In particular, $(\tilde{P}_\varepsilon)$ has a solution.*

Proof of Proposition 4. Let p_ε, $\tilde{p}_\varepsilon$ be the optimal values for (P_ε) and $(\tilde{P}_\varepsilon)$ respectively, where $\tilde{p}_\varepsilon$ is finite as $(\tilde{P}_\varepsilon)$ is feasible and $A_\varepsilon \geq 0$. On one hand, if $\bar{u}_\varepsilon$ is an optimal solution to (P_ε), which exists by Proposition 3, then $(\bar{u}_\varepsilon, T_D\bar{u}_\varepsilon)$ is feasible for $(\tilde{P}_\varepsilon)$. Hence, $\tilde{p}_\varepsilon \leq p_\varepsilon$. On the other, if $(\bar{u}, v)$ feasible for $(\tilde{P}_\varepsilon)$ then $\bar{u}$ is feasible for (P_ε). Since A_ε is monotone, $p_\varepsilon \leq A_\varepsilon(\bar{u}) \leq A_\varepsilon(v)$ whence $p_\varepsilon \leq \tilde{p}_\varepsilon$. We conclude $\tilde{p}_\varepsilon = p_\varepsilon$ and $(\bar{u}_\varepsilon, T_D\bar{u}_\varepsilon)$ is an optimal solution to $(\tilde{P}_\varepsilon)$. □

Under suitable assumptions, the objective function in $(\tilde{P}_\varepsilon)$ is the *difference of convex functions*.

Assumption 3. In addition to Assumptions 1 and 2, there exists $\varphi^+ : \mathbb{R} \to \mathbb{R}$ absolutely continuous, non-decreasing and such that $\varphi^+(t) = 0$ for $t < 0$ and that $\varphi^- := \varphi^+ - \varphi$ is a non-decreasing function.

We let $\varphi^+_\varepsilon(x) = \varphi^+(x/\varepsilon)/\varepsilon$ and we define

$$\Phi^+_\varepsilon(t) = \int_{-\infty}^{t} \varphi^+_\varepsilon(s)ds.$$

Similarly, let $\varphi^-_\varepsilon = \varphi - \varphi^+_\varepsilon$ and $\Phi^-_\varepsilon(t) = \Phi^+_\varepsilon(t) - \Phi_\varepsilon(t)$. By construction, $\Phi_\varepsilon = \Phi^+_\varepsilon - \Phi^-_\varepsilon$. Hence, we can decompose A as $A = A^+_\varepsilon - A^-_\varepsilon$ where

$$A^+_\varepsilon(v) := \int_\Omega \Phi_+(v(x))\, d\mu(x)$$

and A^-_ε is defined similarly.

Proposition 5. *Under Assumption 3, the functionals $A^+_\varepsilon, A^-_\varepsilon : L^\infty_\mu(\Omega) \to \mathbb{R}$ are convex and continuous.*

Proof of Proposition 5. By construction, both Φ^+_ε and Φ^-_ε have derivatives φ^+_ε and φ^-_ε respectively, almost everywhere which are both non-decreasing, hence monotone [88, Proposition 17.10]. Therefore Φ^+_ε and Φ^-_ε are convex. With this, and the linearity and monotonicity of the integral, A^+_ε and A^-_ε are convex. Moreover, they are continuous as the proof of Proposition 3 holds *mutatis mutandis*. □

We conclude that $(\tilde{P}_\varepsilon)$ is a Difference-of-Convex (DC) program [89], [90]. The Convex-Concave Procedure (CCCP) [91] is an efficient method to attempt to find a solution to this class of optimization problems. The CCCP is an iterative method that uses an affine majorant for the concave part, e.g., using subgradients, to majorize the objective function in (15) by a convex function.

Let $v_0 \in L^\infty_\mu(\Omega)$ and let $L^\infty_\mu(\Omega)^*$ be the topological dual of $L^\infty_\mu(\Omega)$. We say $g \in L^\infty_\mu(\Omega)^*$ is a *subgradient* of A^-_ε at v_0 if

$$\forall\, v \in L^\infty_\mu(\Omega) :\ A^-_\varepsilon(v) \geq A^-_\varepsilon(v_0) + g(v - v_0).$$

The *subdifferential* $\partial A^-_\varepsilon(v_0)$ is the collection of all subgradients at v_0. Since $A^-_\varepsilon : L^\infty_\mu(\Omega) \to \mathbb{R}$ is continuous and convex, its subdifferential is non-empty at v_0 [92, Proposition 2.36]. We can use the convex majorizer

$$A_\varepsilon(v) = A^+_\varepsilon(v) - A^-_\varepsilon(v) \leq A^+_\varepsilon(v) - A^-_\varepsilon(v_0) - g_{v_0}(v - v_0)$$

where $g_{v_0} \in L^\infty_\mu(\Omega)^*$ and solve

$$(\tilde{P}_{\varepsilon,v_0}) \quad \begin{cases} \underset{\substack{u\in W_S\cap\mathcal{P} \\ v\in L^\infty_\mu(\Omega)}}{\text{minimize}} & A^+_\varepsilon(v) - A^-_\varepsilon(v_0) - g_{v_0}(v - v_0) \\ \text{subject to} & T_Du \leq v. \end{cases}$$

Proposition 6. *Under Assumption 3, for $\varepsilon > 0$ and $v_0 \in L^\infty_\mu(\Omega)$ $(\tilde{P}_{\varepsilon,v_0})$ has at least one optimal solution.*

We defer the proof to Appendix A-I. In our method we make an explicit choice of a subgradient.

Proposition 7. *Under Assumption 3, for $v_0 \in L^\infty_\mu(\Omega)$ the linear application*

$$g_{v_0}(v) := \int_\Omega \varphi^-_\varepsilon(v_0(x))v(x)\, d\mu(x) \tag{16}$$

is well-defined for $v \in L^\infty_\mu(\Omega)$, continuous and is a subgradient for A^-_ε at v_0.

Proof of Proposition 7. Since $v_0 \in L^\infty_\mu(\Omega)$ and φ^-_ε is non-decreasing, we have $\varphi^-_\varepsilon \circ v_0 \in L^\infty_\mu(\Omega)$ and, as Ω is compact, we also have $\varphi^-_\varepsilon \circ v_0 \in L^1_\mu(\Omega)$. Hence, g_{v_0} is well-defined and $g_{v_0} \in L^\infty_\mu(\Omega)^*$. By the definition of Φ^-_ε and Assumption 3, $\Phi^-_\varepsilon(t) \geq \Phi^-_\varepsilon(t_0) + \varphi^-_\varepsilon(t_0)(t - t_0)$ for all $t, t_0 \in \mathbb{R}$. This implies

$$\int_\Omega \Phi^-_\varepsilon(v(x))\, d\mu(x) \geq \int_\Omega \Phi^-_\varepsilon(v_0(x))\, d\mu(x) + \int_\Omega \varphi^-_\varepsilon(v_0(x))(v(x) - v_0(x))\, d\mu(x).$$

□

Given $(\bar{u}_0, v_0)$ the CCCP constructs a sequence $\{(\bar{u}_k, v_k)\}_{k\in\mathbb{N}}$ where $(\bar{u}_{k+1}, v_{k+1})$ is the optimal solution to $(\tilde{P}_{\varepsilon,v_k})$. To our knowledge, the best theoretical guarantees for finite-dimensional problems show that this sequence converges to a stationary point of $(\tilde{P}_\varepsilon)$ [89, Theorem 3], whereas we are not aware of similar guarantees for infinite-dimensional problems. However, $\{\bar{u}_k\}_{k\in\mathbb{N}}$ is a sequence in

W_S and, as W_S is compact by Proposition 1, we can extract a subsequence $\{\bar{u}_{k(\ell)}\}_{\ell\in\mathbb{N}}$ with limit $\bar{u}^\star$. If $\{\tilde{p}_k^\star\}_{k\in\mathbb{N}}$ is the sequence of optimal values to each $(\tilde{P}_{\varepsilon,v_k})$ then

$$\begin{aligned} A_\varepsilon(T_D\bar{u}^\star) &= \liminf_{\ell\to\infty} A_\varepsilon(T_D\bar{u}_{k(\ell)}) \\ &\le \liminf_{\ell\to\infty} A_\varepsilon(v_{k(\ell)}) \le \liminf_{k\to\infty} \tilde{p}_k^\star \end{aligned}$$

by the continuity of T_D, A_ε and the fact that A_ε is non-decreasing. The optimal values are a conservative estimate of $A_\varepsilon(T_D\bar{u}^\star)$. Our numerical results show the solutions found this way performs well in practice. By iteratively solving $(\tilde{P}_\varepsilon)$ with CCCP for increasingly smaller values of ε we expect to obtain an increasingly accurate approximation to a solution to (P_0). We call this general method Sparse WEighted Error iTeration (SWEET). As we will see, SWEET-ReLU is an specific instance of it.

E. SWEET-ReLU

When φ is the indicator function of $[0,1]$ the function Φ becomes the difference of two *Rectified Linear Units* (ReLUs). In this case, $\varphi_\varepsilon = \varepsilon^{-1}\chi_{[0,\varepsilon]}$. The decomposition $\Phi_\varepsilon = \Phi_\varepsilon^+ - \Phi_\varepsilon^-$ becomes

$$\Phi_\varepsilon^+(x) = x_+/\varepsilon \quad \text{and} \quad \Phi_\varepsilon^-(x) = (x-\varepsilon)_+/\varepsilon$$

and the subgradient (16) becomes

$$g_{v_0}(v) = \frac{1}{\varepsilon}\int_{\{x\in\Omega: v_0(x)>\varepsilon\}} v(x)\,d\mu(x).$$

Let $\Omega_{\varepsilon,v_0} := \{x\in\Omega : v_0(x)\le\varepsilon\}$. Since both $A_\varepsilon^-(v_0)$ and $g_{v_0}(v_0)$ in $(\tilde{P}_{\varepsilon,v_0})$ are constant, it suffices to compute

$$\begin{aligned} A_\varepsilon^+(v) - g_{v_0}(v) &= \frac{1}{\varepsilon}\int_\Omega v(x)_+\,d\mu(x) - \frac{1}{\varepsilon}\int_{\Omega^c_{\varepsilon,v_0}} v(x)\,d\mu(x) \\ &= \frac{1}{\varepsilon}\int_{\Omega_{\varepsilon,v_0}} v(x)_+\,d\mu(x) \\ &\quad + \frac{1}{\varepsilon}\int_{\Omega^c_{\varepsilon,v_0}} (-v(x))_+\,d\mu(x) \end{aligned}$$

where we used the fact that $t_+ - t = (-t)_+$. The second term is non-negative, and becomes positive only when v takes negative values. As $T_D\bar{u}\le v$ in $(\tilde{P}_{\varepsilon,v_0})$ we can choose v arbitrarily large on $\Omega^c_{\varepsilon,v_0}$ to decrease the objective value and to neglect the second integral. Then, only the first term contributes to the objective in $(\tilde{P}_{\varepsilon,v_0})$ and we obtain

$$(\tilde{P}_{\varepsilon,v_0}) \quad \begin{cases} \underset{\substack{\bar{u}\in W_S\cap\mathcal{P}\\ v\in L^\infty_\mu(\Omega)}}{\text{minimize}} & \displaystyle\int_{\Omega_{\varepsilon,v_0}} v(x)_+\,d\mu(x) \\ \text{subject to} & T_D\bar{u}\le v,\ 0\le v|_{\Omega^c_{\varepsilon,v_0}}. \end{cases}$$

As the positive-part function is monotone, we can eliminate v to obtain

$$(\tilde{P}_{\varepsilon,v_0}) \quad \underset{\bar{u}\in W_S\cap\mathcal{P}}{\text{minimize}} \quad \int_{\Omega_{\varepsilon,v_0}} (T_D\bar{u}(x))_+\,d\mu(x)$$

which is precisely the problem (P_1^{SReLU}) when $v_0\equiv 0$. It depends on v_0 only through Ω_{ε,v_0}. To construct an optimal solution $(\bar{u}_k, v_k)$ from the optimal solution $\bar{u}_k$ we proceed as follows: by choosing $v_k|_{\Omega_{\varepsilon,v_{k-1}}} = T_D\bar{u}_k|_{\Omega_{\varepsilon,v_{k-1}}}$ and $v_k|_{\Omega^c_{\varepsilon,v_{k-1}}} = \max\{\varepsilon, T_D\bar{u}_k|_{\Omega^c_{\varepsilon,v_{k-1}}}\}$ we obtain

$$\begin{aligned} \Omega_{\varepsilon,v_k} &= \{x\in\Omega : v_k'(x)\le\varepsilon\} = \{x\in\Omega_{\varepsilon,v_{k-1}} : T_D\bar{u}_k'(x)\le\varepsilon\} \\ &= \Omega_{\varepsilon,v_{k-1}}\cap\{x\in\Omega : T_D\bar{u}_k'(x)\le\varepsilon\}, \end{aligned}$$

yielding the method as presented in Section III.

F. A class of monaural dissimilarity maps

We introduce a class of dissimilarity metrics based on time-variant filters satisfying Assumption 1.

Lemma 1. *Let $K : Z\to L^2(\mathbb{R}^2)$ be continuous with*

$$\sup_{(x,\theta)\in Z}\sup_{t\in\mathbb{R}}\int_{\mathbb{R}} (|K_{(x,\theta)}(t,t')| + |K_{(x,\theta)}(t',t)|)\,dt'$$

finite. Define for $(x,\theta)\in Z$, $w\in L^2(\mathbb{R})$

$$A^K_{(x,\theta)}w(t) := \int_{\mathbb{R}} K_{(x,\theta)}(t,t')w(t')\,dt'.$$

Then $A^K_{(x,\theta)}$ is linear, $A^K_{(x,\theta)}w\in L^2(\mathbb{R})$ and $(x,\theta,w)\mapsto A^K_{(x,\theta)}w$ is continuous.

Proof. The linearity follows from the integral representation. From Young's inequality for integral operators [93, Theorem 0.3.1]

$$\int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{(x,\theta)}(t,t')w(t')\,dt'\right|^2 dt \le C_K(x,\theta)^2\|w\|_{L^2}^2$$

whence $(x,\theta)\mapsto A_{(x,\theta)}w$ is bounded. By the Cauchy-Schwarz inequality

$$\begin{aligned} &\|A_{(x,\theta)}w' - A_{(y,\phi)}w\|_{L^2}^2 \\ &\quad\le 2\int_{\mathbb{R}}\left|\int_{\mathbb{R}}(K_{(x,\theta)}(t,t') - K_{(y,\phi)}(t,t'))w'(t')\,dt'\right|^2 dt \\ &\qquad + 2\int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{(y,\phi)}(t,t')(w'(t') - w(t'))\,dt'\right|^2 dt \\ &\quad\le 2\|w\|_{L^2}^2\iint_{\mathbb{R}^2}|K_{(x,\theta)}(t,t') - K_{(y,\phi)}(t,t')|^2 dt dt' \\ &\qquad + 2C_K(x,\theta)\|w' - w\|_{L^2}^2, \end{aligned}$$

where continuity follows from hypothesis. □

Proposition 8. *Let $\bar{u}_0\in W$ and let $\{B_k\}_{k=1}^{n_b}$ be as in* (10) *where $\{K_{B_k}\}_{k=1}^{n_b}$ satisfy the hypotheses of Lemma 1. Let $\Psi : \mathbb{R}_+^{n_b}\to\mathbb{R}$ be convex and monotone increasing on each one of its arguments. For $s\in\{\ell, r\}$*

$$\begin{aligned} D^s_{(\bar{u},\bar{u}_0)}(x,\theta) = \Psi(&B_1(u^s_{(x,\theta)} - u^s_{0,(x,\theta)}), \\ &\ldots, B_{n_b}(u^s_{(x,\theta)} - u^s_{0,(x,\theta)})) \end{aligned}$$

is a dissimilarity satisfying Assumption 1. In particular, so is

$$D_{(\bar{u},\bar{u}_0)}(x,\theta) = \max\{D^\ell_{(\bar{u},\bar{u}_0)}(x,\theta), D^r_{(\bar{u},\bar{u}_0)}(x,\theta)\}.$$

Proof. For simplicity, we prove the result for $\bar{u}_0 = 0$. It can be verified that $B_k u^s_{(x,\theta)} = \|A^{K_{B_k}}_{(x,\theta)}u^s_{(x,\theta)}\|_{L^2}^2$. By Lemma 1, a function of the form

$$D^s_{(\bar{u},0)}(x,\theta) = \Psi(\|A^{K_{B_1}}_{(x,\theta)}u^s_{(x,\theta)}\|_{L^2},\ldots,\|A^{K_{B_{n_b}}}_{(x,\theta)}u^s_{(x,\theta)}\|_{L^2})$$

is continuous on Z. Since $A^{K_{B_k}}_{(x,\theta)}$ is linear on $u^s_{(x,\theta)}$ and the norm is convex, the convexity of D^s follows from the assumptions on Ψ. □

G. Discussion

Although Proposition 2 implies A_ε converges pointwise to $\mu \circ \mathcal{S}$, this is not sufficient to ensure a global minimizer for (P_ε) converges to a global minimizer for (P_0). A future line of work consists on leveraging Γ-convergence to answer this question. Related to it is the choice of φ. Although we have not studied extensively the effect of this choice, we believe it affects the quality of the approximation to the area of the sweet spot. This is another interesting future line of work. Finally, the method allows for several choices of μ. Therefore, the results presented apply both for the continuous case, e.g., when μ is the Lebesgue measure, and the discrete case, e.g., when μ is discrete.

H. Proof of Proposition 1

Proof of (i). It is easy to verify W_S is convex. Since Z is a separable metric space, by Arzelà-Ascoli's theorem [94, Theorem 11.28] to prove W_S is compact it suffices to show it is bounded and equicontinuous. By Assumption 1,

$$\begin{aligned}\|u^s_{(x,\theta)}\|^2_{L^2} &\leq n_s \sum_{k=1}^{n_s} \int_{I_S} |\widehat{\alpha}_k(f)|^2 |H^s_k(f,x,\theta)|^2 \, df \\ &\leq n_s \gamma^2_{\max} \sum_{k=1}^{n_s} \sup_{(f,x,\theta)\in I_S \times Z} |H^s_k(f,x,\theta)|^2\end{aligned}$$

and $(x,\theta) \mapsto u_{(x,\theta)}$ is uniformly bounded. Thus, W_S is bounded. Let $\varepsilon > 0$. Since $\widehat{H}^s_k$ is continuous on the compact set $I_S \times Z$, there is $\delta > 0$ such that for any $|x-y|, |\theta - \phi| < \delta$ and $f \in I_S$ we have that $|\widehat{H}^s_k(f,x,\theta) - \widehat{H}^s_k(f,y,\phi)| < \varepsilon / 2 n_s^2 \gamma^2_{\max}$. Then,

$$\begin{aligned}&\|u^s_{(x,\theta)} - u^s_{(y,\phi)}\|^2_{L^2} \\ &\leq n_s \sum_{k=1}^{n_s} \int_{I_S} |\widehat{\alpha}_k(f)|^2 |H^s_k(f,x,\theta) - H^s_k(f,y,\phi)|^2 \, df < \frac{1}{2}\varepsilon\end{aligned}$$

whence $(x,\theta) \mapsto \bar{u}_{(x,\theta)}$ is continuous. Since δ is independent of $\bar{u}$, we conclude W_S is equicontinuous.

Proof of (ii). Let $\bar{u}, \bar{u}_0 \in W$. The function $D_{(\bar{u},\bar{u}_0)}$ is continuous on the compact set Z and thus bounded. Hence, $T_D \bar{u}$ is bounded and we can associate to it its equivalence class in $L^\infty_\mu(\Omega)$. Let $\varepsilon > 0$. There exists $\delta > 0$ such that $\|v^\ell - u^\ell\|_{L^2}, \|v^r - u^r\|_{L^2} < \delta$ implies $|D_{(\bar{v},\bar{u}_0)}(x,\theta) - D_{(\bar{u},\bar{u}_0)}(x,\theta)| < \varepsilon/2$. This implies $|T_D \bar{v}(x) - T_D \bar{u}(x)| < \varepsilon$ whence T_D is continuous. By Assumption 1, the map D is convex on its first argument. The conclusion follows from the fact that the pointwise supremum of convex functions is convex.

Proof of (iii). For $T_D u \in L^\infty_\mu(\Omega)$ we can choose representatives $T_D u', T_D u''$ of $T_D u$. Hence, the set $\{x \in \Omega : T_D u'(x) = T_D u''(x)\}$ has μ-measure zero, from where the conclusion follows.

Proof of (iv). From the same arguments used for (ii) the map $T_L : W \to L^\infty_\mu(\Omega)$ is continuous. As $\{v \in L^\infty_\mu(\Omega) : vv > 0 \ \mu\text{-a.e.}\}$ is open, then $\mathcal{P}^c = \{\bar{u} \in W : T_L \bar{u} > 0 \ \mu\text{-a.e.}\}$ is open, whence $\mathcal{P}$ is closed.

I. Proof of Proposition 6

Let v'_0 be a representative and define

$$f(x,\alpha) = \Phi^+_\varepsilon(\alpha) - \varphi^+_\varepsilon(v'_0(x))\alpha.$$

This is a Carathéodory map [95, Definition 8.2.7]. By Theorem 8.2.11 in [95] there exists $v'^\star$ measurable such that

$$\Phi^+_\varepsilon(v'^\star(x)) - \varphi^+_\varepsilon(v'_0(x))v'^\star(x) = \inf\{f(x,\alpha) : \alpha \in \mathbb{R}\}.$$

By Assumption 3, $\varphi^+_\varepsilon(v'_0(x)) \geq 0$ and $0 \leq v'^\star(x) \leq v'_0(x)$ whence $v'^\star$ is bounded μ-a.e. Let $v^\star \in L^\infty_\mu(\Omega)$ be its equivalence class. Let $\{(\bar{u}_k, v_k)\}_{n\in\mathbb{N}}$ be a minimizing sequence. As $W_S \cap \mathcal{P}$ is compact, without loss of generality we may assume $\{\bar{u}_k\}_{k\in\mathbb{N}}$ has a limit $\bar{u}_\infty$. Let $T_D \bar{u}'_k$ be a representative and let $w'_k := \max(v'^\star, T_D u'_k)$. Then w'_k is μ-a.e. bounded. Let $w_k \in L^\infty_\mu(\Omega)$ denote its equivalence class. By construction,

$$f(x, v'_k(x)) \geq f(x, T_D \bar{u}'_k(x)) \geq f(x, T_D v'^\star(x))$$

for any representative v'_k. Therefore

$$\liminf_{k\to\infty}(A^+_\varepsilon(v_k) - g_{v_0}(v_k)) \geq \liminf_{k\to\infty}(A^+_\varepsilon(w_k) - g_{v_0}(w_k))$$

whence $\{(\bar{u}_k, w_k)\}_{k\in\mathbb{N}}$ is also minimizing. By continuity of T_D we conclude $w_k \to \max\{v^\star, T_D \bar{u}_\infty\}$ whence $(\tilde{P}_{\varepsilon,v_0})$ has a solution.

Acknowledgment

We thank the anonymous referees for their comments and Julius O. Smith for insightful discussions about the topic. The authors were partially funded by a ArTeCiH grant, Dirección de Artes y Cultura, VRI-UC. C. A. SL. was partially funded by ANID – FONDECYT – 1211643, ANID – Millennium Science Initiative Program – NCN17_059 and ANID – Millennium Science Initiative Program – NCN17_129.

References

[1] R. Nicol, "Creating auditory illusions with spatial-audio technologies," in *The Technology of Binaural Understanding*. Springer, 2020, pp. 581–622.

[2] H. Wierstorf, A. Raake, and S. Spors, "Binaural assessment of multichannel reproduction," in *The technology of binaural listening*. Springer, 2013, pp. 255–278.

[3] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[4] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[5] D. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959.

[6] H. Wierstorf, "Perceptual assessment of sound field synthesis," Ph.D. dissertation, Technische Universitaet Berlin (Germany), 2014.

[7] M. Frank and F. Zotter, "Exploring the perceptual sweet area in ambisonics," *Journal of the Audio Engineering Society*, may 2017.

[8] F. Rumsey, "Surround sound," in *In Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. Focal Press, 2017, p. Chap. 6.

[9] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering sSciety*, vol. 45, no. 6, pp. 456–466, 1997.

[10] ——, "Coloration of amplitude-panned virtual sources," in *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

[11] C. Huygens, *Traité de la lumière*. Pierre Vander Aa Marchand Libraire, 1690.

[12] 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, University of Paris VI, 2000.

[13] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[14] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

[15] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 9, 2001.

[16] N. Ueno, S. Koyama, and H. Saruwatari, "Three-dimensional sound field reproduction based on weighted mode-matching method," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, pp. 1852–1867, 12 2019.

[17] H. Zuo, T. D. Abhayapala, and P. N. Samarasinghe, "Particle velocity assisted three dimensional sound field reproduction using a modal-domain approach," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 2119–2133, 2020.

[18] ——, "3d multizone soundfield reproduction in a reverberant environment using intensity matching method," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 416–420.

[19] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992–3000, 1993.

[20] G. N. Lilis, D. Angelosante, and G. B. Giannakis, "Sound field reproduction using the lasso," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 1902–1912, 2010.

[21] M. Kolundzija, C. Faller, and M. Vetterli, "Sound field reconstruction: An improved approach for wave field synthesis," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[22] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.

[23] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[24] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *In 124th Convention of the AES*. Citeseer, 2008.

[25] H. Wierstorf, A. Raake, and S. Spors, "Assessing localization accuracy in sound field synthesis," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 1111–1119, 2017.

[26] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake, "Coloration in wave field synthesis," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society, 2014.

[27] S. Spors and J. Ahrens, "A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling," in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[28] F. M. Fazi and P. A. Nelson, "A theoretical study of sound field reconstruction techniques," in *19th International Congress on Acoustics*, September 2007.

[29] J. D. Johnston and Y. H. V. Lam, "Perceptual soundfield reconstruction," in *Audio Engineering Society Convention 109*. Audio Engineering Society, 2000.

[30] E. D. Sena, H. Hacihabiboglu, and Z. Cvetkovic, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 1653–1665, 2013.

[31] T. Ziemer and R. Bader, "Psychoacoustic sound field synthesis for musical instrument radiation characteristics," *AES: Journal of the Audio Engineering Society*, vol. 65, pp. 482–496, 6 2017.

[32] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[33] T. Lee, J. K. Nielsen, and M. G. Christensen, "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 2412–2426, 2020.

[34] L. C. Evans, *Partial differential equations*, 2nd ed. American Mathematical Society, 2010.

[35] P.-A. Gauthier, A. Berry, and W. Woszczyk, "Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain," *The Journal of the Acoustical Society of America*, vol. 117, pp. 662–678, 2 2005.

[36] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[37] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[38] S. G. Krantz and H. R. Parks, *A primer of real analytic functions*. Boston, MA: Birkhäuser Boston, 2002.

[39] D. Deutsch, "Auditory illusions, handedness, and the spatial environment," *Journal of the Audio Engineering Society*, vol. 31, no. 9, pp. 606–620, september 1983.

[40] J. Francombe, T. Brookes, and R. Mason, "Elicitation of the differences between real and reproduced audio," *Journal of the Audio Engineering Society*, may 2015.

[41] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[42] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[43] J. Blauert, "Models of binaural hearing: architectural considerations," in *Proc. 18 th DANAVOX Symposium 1999*, 1999, pp. 189–206.

[44] A. Raake and H. Wierstorf, "Binaural evaluation of sound quality and quality of experience," in *The Technology of Binaural Understanding*. Springer, 2020, pp. 393–434.

[45] A. Raake and J. Blauert, "Comprehensive modeling of the formation process of sound-quality," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 76–81.

[46] T. Letowski, "Sound quality assessment: concepts and criteria," in *Audio Engineering Society Convention 87*. Audio Engineering Society, 1989.

[47] J. Francombe, T. Brookes, and R. Mason, "Evaluation of spatial audio reproduction methods (part 1): Elicitation of perceptual differences," *Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 198–211, march 2017.

[48] M. Dietz and G. Ashida, "Computational models of binaural processing," in *Binaural Hearing*. Springer, 2021, pp. 281–315.

[49] S. Raleigh Lord, "On our perception of sound direction," *Phil Mag*, vol. 13, pp. 314–232, 1907.

[50] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[51] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, 2011.

[52] A. Brughera, L. Dunai, and W. M. Hartmann, "Human interaural time difference thresholds for sine tones: The high-frequency limit," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2839–2855, 2013.

[53] E. C. Cherry and B. M. A. Sayers, ""human 'cross-correlator'"—a technique for measuring certain parameters of speech perception," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 889–895, 1956.

[54] L. A. Jeffress, "A place theory of sound localization," *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[55] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1608–1622, 1986.

[56] V. Pulkki and T. Hirvonen, "Functional count-comparison model for binaural decoding," *Acta Acustica united with Acustica*, vol. 95, no. 5, pp. 883–900, 2009.

[57] D. McAlpine and B. Grothe, "Sound localization and delay lines–do mammals fit the model?" *Trends in neurosciences*, vol. 26, no. 7, pp. 347–350, 2003.

[58] J. F. Culling and M. Lavandier, "Binaural unmasking and spatial release from masking," in *Binaural Hearing*. Springer, 2021, pp. 209–241.

[59] M. Brüggen, "Sound coloration due to reflections and its auditory and instrumental compensation," Ph.D. dissertation, PhD thesis, Ruhr-Universität Bochum, 2001.

[60] J. Breebaart, S. Van De Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. i. model structure," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1074–1088, 2001.

[61] M. Park, P. A. Nelson, and K. Kang, "A model of sound localisation applied to the evaluation of systems for stereophony," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 825–839, 2008.

[62] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing virtual sound source attributes using a binaural auditory model," *Journal of the Audio Engineering Society*, vol. 47, no. 4, pp. 203–217, 1999.

[63] T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney, "Predicting the colouration between binaural signals," *Applied Sciences*, vol. 12, no. 5, p. 2441, 2022.

[64] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[65] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1323–1331, 1993.

[66] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1–13, 2005.

[67] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[68] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America*, vol. 124, pp. 422–438, 7 2008.

[69] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 310–319, 1 2007.

[70] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 1553–1564, 2012.

[71] K. A. B. Knobel and T. G. Sanchez, "Nível de desconforto para sensação de intensidade em indivíduos com audição normal," *Pró-Fono Revista de Atualização Científica*, vol. 18, no. 1, pp. 31–40, 2006.

[72] L. P. Sherlock and C. Formby, "Estimates of loudness, loudness discomfort, and the auditory dynamic range: normative estimates, comparison of procedures, and test-retest reliability," *Journal of the American Academy of Audiology*, vol. 16, no. 02, pp. 085–100, 2005.

[73] E. Terhardt, "Calculating virtual pitch," *Hearing research*, vol. 1, no. 2, pp. 155–182, 1979.

[74] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[75] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*. Springer Science & Business Media, 2010, vol. 37.

[76] P. Izquierdo Lehmann, R. Cádiz, and C. A. Sing Long, "WFR Tools." [Online]. Available: https://github.com/csl-lab/sweet-sps

[77] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[78] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.

[79] M. ApS, *The MOSEK optimization toolbox for Python manual. Version 9.2.44*, 2019.

[80] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," in *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

[81] J. Ahrens and S. Spors, "Spatial encoding and decoding of focused virtual sound sources," in *Ambisonics Symposium*, 2009, pp. 25–27.

[82] J. Ahrens, *Analytic methods of sound field synthesis*. Springer Science & Business Media, 2012.

[83] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[84] P. Majdak, C. Hollomey, and R. Baumgartner, "Amt 1.0: The toolbox for reproducible research in auditory modeling," *submitted to Acta Acustica*, 2021.

[85] J. Ahrens and S. Spors, "Focusing of virtual sound sources in higher order ambisonics," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[86] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–976, 2005.

[87] D. L. Cohn, *Measure theory*, 2nd ed., ser. Birkhäuser Advanced Texts Basler Lehrbücher. New York, NY: Springer New York, 2013.

[88] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, ser. CMS Books in Mathematics. New York, NY: Springer New York, 2011.

[89] P. D. Tao and L. T. H. An, "Convex analysis approach to D.C. programming: Theory, algorithms and applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[90] R. Horst and N. V. Thoai, "DC programming: Overview," *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, oct 1999.

[91] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, jun 2016.

[92] V. Barbu and T. Precupanu, *Convexity and optimization in Banach spaces*, 4th ed., ser. Springer Monographs in Mathematics. Dordrecht: Springer Netherlands, 2012.

[93] C. D. Sogge, *Fourier integrals in classical analysis*, 2nd ed. Cambridge: Cambridge University Press, 2017.

[94] W. Rudin, *Real and complex analysis*, 3rd ed. McGraw-Hill Education, 1986.

[95] J.-P. Aubin and H. Frankowska, *Set-valued analysis*. Birkhauser, 1990.

Pedro Izquierdo Lehmann (S'21) is currently finishing the M.Sc. degree in Applied Mathematics over the field of Signal Processing from Pontificia Universidad Católica de Chile. Also he works as a Research Assistant at the Biomedical Institute of the same university. His research interests include mathematical optimization, signal processing and statistical learning, with an emphasis on spatial sound and biomedical applications.

Rodrigo F. Cádiz (Member, IEEE) holds B.S. and B.A. degrees in Electrical Engineering and Music Composition from the Pontificia Universidad Católica de Chile (UC). He obtained his Ph.D. in Music Technology from Northwestern University in 2006. His research interests include sound synthesis, digital signal processing, computer music, composition, sonification, new interfaces for musical expression and the musical applications of artificial intelligence and complex systems. He is currently full professor at the Music Institute and Electrical Engineering Department of UC.

Carlos A. Sing-Long (S'12–M'16) received the Diplôme d'Ingénieur from Ecole Polytechnique in 2007 and both a professional degree in Electrical Engineering and a B.S. degree in Physics from Pontificia Universidad Católica de Chile in 2008. In 2016, he received a M.S. and a Ph.D. degree in Computational and Mathematical Engineering from Stanford University. Since 2016, he has been Assistant Professor at Pontificia Universidad Catolica de Chile with a joint appointment at the Institute for Mathematical and Computational Engineering and the Institute for Biological and Medical Engineering. His research focus is on the mathematical analysis of discrete inverse problems with an emphasis on imaging applications. His research interests include statistical signal processing, mathematical imaging, applied harmonic analysis, and mathematical optimization.