Proceedings of Meetings on Acoustics

Volume 30 http://acousticalsociety.org/

Acoustics `17 Boston

173rd Meeting of Acoustical Society of America and 8th Forum Acusticum
Boston, Massachusetts
25-29 June 2017

Psychological and Physiological Acoustics: Paper 1pPPb4

Models for evaluating navigational techniques for higher-order ambisonics

Joseph G. Tylka and Edgar Y. Choueiri
Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey, 08544, USA; josephgt@princeton.edu; choueiri@princeton.edu

Models are presented that predict perceived source localization and spectral coloration for the purpose of evaluating navigational techniques for higher-order ambisonics. Previous evaluations typically rely on binaural localization models, which conflate the effects of the navigational technique with those of the adopted ambisonics-to-binaural rendering approach. Moreover, studies on navigation-induced coloration have been largely qualitative. The presented models are applied directly to translated ambisonics impulse responses (i.e., before rendering to binaural) and are validated through listening experiments. Localization is predicted using an extension to a precedence-effect-based localization model. Coloration is predicted using a linear combination of spectral energies and notch-depths in a difference-spectrum between the test and reference signals. For two interpolation-based navigational techniques and a range of translation distances, localization and coloration are also measured subjectively through binaural-synthesis-based listening tests, wherein subjects judge source position for a spatialized sample of speech and rate the induced coloration in pink noise relative to reference signals. The proposed localization model is shown to predict the data with comparable accuracy to that of a binaural localization model and the coloration metrics used are shown to best predict perceived coloration compared to alternative sets of metrics.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0000625

1. INTRODUCTION

Virtual navigation of three-dimensional higher-order ambisonics sound fields (i.e., sound fields that have been decomposed into spherical harmonics) enables a listener to explore an acoustic space and experience a spatially-accurate perception of the sound field. Applications of sound field navigation may be found in virtual-reality reproductions of real-world spaces. For example, to reproduce an orchestral performance in virtual reality, navigation of an acoustic recording of the performance may yield superior spatial and tonal fidelity compared to that produced through acoustic simulation of the performance. Navigation of acoustic recordings may also be preferable when reproducing real-world spaces for which computer modeling of complex wave-phenomena and room characteristics may be too computationally intensive for real-time playback and interaction.

Recently, several navigational techniques for higher-order ambisonics have been developed, all of which may degrade localization information and induce spectral coloration. The severity of such penalties needs to be investigated and quantified in order to both compare existing navigational techniques and develop novel ones. Although subjective testing is the most direct method of evaluating and comparing navigational techniques, such tests are often lengthy and costly, which motivates the use of objective metrics that enable quick assessments of navigational techniques.

A. PREVIOUS WORK AND REMAINING PROBLEMS

Several recent studies have investigated localization accuracy of various navigational techniques. Winter et al. (2014) evaluated the localization accuracy of a plane-wave-based translation technique (Schultz and Spors, 2013) using a binaural localization model (Dietz et al., 2011) to predict perceived localization. Tylka and Choueiri (2015) compared the localization errors incurred by various translation techniques[1] using the velocity and energy localization vectors developed by Gerzon (1992). However, this analysis neglected the precedence effect, which is expected to play an important role in the context of sound field navigation, as an accurate virtual translation of the listener necessarily involves direction-dependent time shifting of incident signals. Consequently, more recently, Tylka and Choueiri (2016) evaluated the localization accuracy of a proposed interpolation-based navigational technique using an extension of a precedence-effect-based localization vector developed by Stitt et al. (2016), which was itself an extension of the original energy vector. Although the model of Stitt et al. has been validated through listening tests and shows improvements compared to binaural localization models (specifically, those by Dietz et al. (2011); Lindemann (1986)) the more recent extension by Tylka and Choueiri (2016) has not been validated through listening tests.

Additionally, recent studies have investigated spectral colorations induced by various navigational techniques. Hahn and Spors (2015) evaluated spectral coloration induced by the plane-wave translation technique (Schultz and Spors, 2013) by visually examining impulse and frequency responses. In a similar manner, Tylka and Choueiri (2016) evaluated and compared the spectral coloration induced by their proposed interpolation technique and the linear interpolation technique of Southern et al. (2009). While it is clear from these studies that most (if not all) existing navigational techniques tend to induce at least some spectral coloration, both analyses were largely qualitative. Consequently, it remains difficult to compare these colorations between techniques without numerical measures of perceptible coloration.

B. OBJECTIVES AND APPROACH

Here, we present models for perceived source localization and spectral coloration which we have developed for the purpose of evaluating and comparing techniques for sound field navigation. In order to isolate

[1] The techniques compared were: virtual ambisonics playback (see Tylka and Choueiri, 2015, section 3.1), plane-wave expansion and translation (Schultz and Spors, 2013), and spherical-harmonic re-expansion (Gumerov and Duraiswami, 2005, chapter 3).

any errors introduced by the navigational technique under test (which operates in the ambisonics domain) from those introduced through rendering to binaural, we sought models that are independent of any choice of ambisonics-to-binaural rendering approach.[2] Consequently, we developed the proposed models such that they operate directly on ambisonics impulse responses. In contrast, although not explicitly investigated here, we expect that the predictions of binaural models may be sensitive to the choice of binaural rendering approach and/or to the choice of head-related transfer function (HRTF).

We also sought models that are perceptually relevant, in that the predictions of the models should agree with subjective listening test responses. Consequently, we conducted two listening experiments (one each for localization and coloration), the results of which were used to determine parameters of the proposed models. We then validated the proposed models through comparisons against alternative models in terms of their agreement with the experimental data.

In Sec. 2, we present the localization and coloration models developed in this work and, in Sec. 3, we describe the corresponding listening experiments. In Sec. 4, we compare the results of the listening experiments with predictions of the models and, in Sec. 5, we conclude and discuss avenues for future work.

2. AUDITORY MODELS

In this section, we describe the models used for predicting perceived localization and spectral coloration.

A. LOCALIZATION MODELS

Here, we describe two localization models, each of which requires only an ambisonics impulse response (either before or after navigation) and a stimulus signal (e.g., pink noise, speech, music, etc.) in order to predict localization.

i. Perceptually-Weighted Localization Vector

Recently, Stitt et al. (2016) proposed an extension to incorporate the precedence effect into the energy vector of Gerzon (1992). In their paper, Stitt et al. also showed that their proposed model achieves improved localization accuracy compared to the binaural models of Dietz et al. (2011) and Lindemann (1986). Motivated by these results, we proposed in a recent paper (Tylka and Choueiri, 2016) an extension to this localization model; here, we discuss and revise our proposed extension.

We begin by converting the ambisonics impulse response, $a_n(t)$, into a set of Q impulse responses for a specified grid of plane-wave directions, $\hat{r}_q$. For ambisonics order L, there are $N = (L+1)^2$ ambisonics signals, and the corresponding plane-wave signals, μ, are given by (Gumerov and Duraiswami, 2005, section 2.3.3)

$$\mu(t, \hat{r}_q) = \sum_{n=0}^{N-1} a_n(t) Y_n(\hat{r}_q), \tag{1}$$

where Y_n is the n^{th} real-valued orthonormal (N3D) spherical harmonic and n is the ambisonics channel number (ACN), as defined by Nachbar et al. (2011). The grids of directions (also called "nodes") used here are given by Fliege and Maier (1999), who also provide the corresponding quadrature weights.[3] As this discrete plane-wave sound field would generally be rendered via quadrature integration (e.g., see Eq. (8)

[2]Examples of such binaural rendering approaches include computing a plane-wave expansion and convolving each term by the corresponding head-related transfer function (Duraiswami et al., 2005) and non-linear, parametric techniques such as HARPEX (Berge and Barrett, 2010).

[3]Node coordinates and corresponding quadrature weights can be found here: `http://www.mathematik.uni-dortmund.de/lsx/research/projects/fliege/nodes/nodes.html`

below), the impulse response for each plane-wave term is given by $w_q \mu(t, \hat{r}_q)$, where w_q is the quadrature weight for the direction $\hat{r}_q$.

Next, we identify and isolate temporally-distinct impulse response "wavelets." To do this, we apply a 4^{th}-order Butterworth high-pass filter with a cut-off frequency of 500 Hz to all impulse responses in the set and compute the global maximum (i.e., the largest absolute value over all impulse responses in the set). For each impulse response, we take the absolute value and identify any peaks (i.e., local maxima) whose amplitudes are at least γ dB relative to the global maximum. If no such peaks exist in a given impulse response, then that response in its entirety is treated as a wavelet. If at least one such peak exists, then, around each peak, we apply a Tukey window beginning τ ms before the peak and ending either τ ms after the peak, or at the position of the following peak, whichever yields a larger window length. Both the cosine fade-in and fade-out of the Tukey window are τ ms in duration. In this way, a single impulse response may be split into several wavelets. For each wavelet, we apply a 10% (-20 dB, now relative to the peak of the wavelet) threshold to determine the time-delay of the onset.

For the purposes of this model, we consider each wavelet to be a distinct sound source, such that wavelets extracted from the same impulse response originate from the same direction, but at different times, given by their onset times. Taking the Fourier transform of each wavelet yields complex-valued, frequency-dependent gains, $G_{q'}$, of the q'^{th} wavelet, where $q' \in [1, Q']$ and $Q' \geq Q$ is the total number of wavelets. We then average these gains in critical bands using a gammatone filter bank.[4] The frequency-averaged gain is given by

$$\overline{G}_{q'}(f_c) = \frac{\int_{-\infty}^{\infty} |\Gamma(f; f_c)||G_{q'}(f)|df}{\int_{-\infty}^{\infty} |\Gamma(f; f_c)|df}, \tag{2}$$

where $\Gamma(f; f_c)$ is a gammatone filter with center frequency f_c for $c \in [1, N_b]$, for a set of ERB-spaced (equivalent rectangular bandwidth) center frequencies (Glasberg and Moore, 1990) spanning the range $f \in [20\text{ Hz}, 20\text{ kHz}]$.

For each frequency band, we feed these gains into the model, yielding a frequency-dependent predicted localization vector $\vec{r}_{\text{PE}}(f_c)$, given by (Stitt et al., 2016)

$$\vec{r}_{\text{PE}}(f_c) = \frac{\sum_{q'} |w_{q'} \overline{G}_{q'}(f_c)|^2 \, \hat{r}_{q'}}{\sum_{q'} |w_{q'} \overline{G}_{q'}(f_c)|^2}. \tag{3}$$

where $w_{q'}$ is a perceptual weight (based on the precedence-effect) for the q'^{th} wavelet. Similarly, we define a perceptually-weighted velocity vector, given by

$$\vec{r}_{\text{PV}}(f_c) = \frac{\sum_{q'} |w_{q'} \overline{G}_{q'}(f_c)| \, \hat{r}_{q'}}{\sum_{q'} |w_{q'} \overline{G}_{q'}(f_c)|}. \tag{4}$$

Note that this definition differs from that for the original velocity vector, given by Gerzon (1992), in which complex-valued gains are used and the real part of the resulting vector is taken. This modification may be justified since the time-dependence of the sources is captured in the precedence-effect-based weights, so the source's phase is no longer needed.

We then combine the velocity vector below 700 Hz and the energy vector above into a single, frequency-dependent vector, given by

$$\vec{r}_{\text{PC}}(f_c) = \begin{cases} \beta \cdot \vec{r}_{\text{PV}}(f_c), & \text{for } f_c \leq f_{\text{XO}}, \\ \vec{r}_{\text{PE}}(f_c), & \text{for } f_c > f_{\text{XO}}, \end{cases} \tag{5}$$

[4]In this work, we used the gammatone filters implemented in the large time-frequency analysis toolbox (LTFAT) for MATLAB: `http://ltfat.sourceforge.net/`

where f_{XO} is the "crossover" frequency, equal to the center frequency nearest to 700 Hz, and β is a normalization factor to match low- and high-frequency vector magnitudes, given by $\beta = \|\vec{r}_{\mathrm{PE}}(f_{\mathrm{XO}})\| \,/\, \|\vec{r}_{\mathrm{PV}}(f_{\mathrm{XO}})\|$, where $\|\cdot\|$ denotes the ℓ^2 norm (Euclidean distance) of a vector. Finally, we compute a weighted-average vector, which depends on the stimulus signal and is given by

$$\vec{r}_{\mathrm{P}} = \frac{\sum_c X_c \, \vec{r}_{\mathrm{PC}}(f_c)}{\sum_c X_c}, \tag{6}$$

where the weights X_c are the stimulus signal's energy in each critical band, given by

$$X_c = \int_{-\infty}^{\infty} |\Gamma(f; f_c)||X(f)|^2 df, \tag{7}$$

and $X(f)$ is the Fourier transform of the stimulus signal.

In addition to the 3 free model parameters (Q, τ, and γ) defined above, the original model of Stitt et al. (2016) retains one free parameter, $\alpha \in [0, 1]$, which specifies the relative importance of stationary (i.e., time-averaged) to transient information in the stimulus signal.[5] As described in Sec. 4.A, we determined optimal values for these 4 parameters based on a best fit of the model's predictions to the data from the listening experiment.

ii. Binaural Localization Model

For comparison, we also predicted localization using the binaural localization model of Dietz et al. (2011). In order to compute the required binaural impulse responses, the ambisonics impulse response is first converted to plane-wave impulse responses using Eq. (1). The binaural impulse responses are then computed by (Duraiswami et al., 2005, Eq. (30))

$$b^{\mathrm{L,R}}(t) = \sum_{q=1}^{Q} w_q \mu(t, \hat{r}_q) * h^{\mathrm{L,R}}(t, \hat{r}_q), \tag{8}$$

where '$*$' denotes convolution, Q is again the number of plane-wave terms, $h^{\mathrm{L,R}}$ denotes the head-related impulse response for source direction $\hat{r}_q$, and the 'L,R' superscripts refer to the left and right ears, respectively.

An implementation of this model is freely available in the auditory modeling toolbox[6] (Søndergaard and Majdak, 2013). In this work, we adopted the extension proposed by Wierstorf et al. (2013), in which an ITD-to-azimuth lookup table is first generated for each subject using that subject's measured HRTFs. The stimulus signal is then filtered by the binaural impulse responses and the ITD is computed using the original model. The model yields ITD in a set of frequency bands, which are then converted to azimuth via the lookup table. Outliers beyond 30° away from the median azimuth are then removed, and finally, a single predicted azimuth is computed as the weighted average over frequency, with weights given by the rms signal amplitude in each frequency band.

B. COLORATION METRICS

In this work, we followed the approach of Wittek et al. (2007) and developed linear regression models to predict subjective ratings of coloration given some combination of the coloration metrics described

[5] For example, a highly transient signal is expected to require a low value of α, while a more stationary signal would require a higher value (Stitt et al., 2016, 2017).

[6] Available here: `http://amtoolbox.sourceforge.net/`

below. As will become clear below, we define each of these metrics *relative* to some reference signal. Consequently, each metric is computed using both a *test sample* (i.e., the HOA impulse response for the listening position after processing through some navigational technique) and a *reference sample* (i.e., the HOA impulse response captured directly at the listening position).

i. Auditory Band Spectral Error (ABSE)

The *auditory band spectral error* (ABSE), adapted from Schärer and Lindau (2009, Eq. (9)), is given by

$$\mathrm{ABSE}(f_c) = 10\log_{10}\left(\frac{\int_{-\infty}^{\infty}|\Gamma(f;f_c)||F_\mathrm{T}(f)|^2 df}{\int_{-\infty}^{\infty}|\Gamma(f;f_c)||F_\mathrm{R}(f)|^2 df}\right), \tag{9}$$

where F_T is the free-field transfer function of the test sample, and F_R is the free-field transfer function of the reference sample. Each free-field transfer function is obtained by taking the Fourier transform of the zeroth-order (i.e., omnidirectional) term of the respective ambisonics impulse response. For this and other metrics requiring an auditory filter bank, we use ERB-spaced center frequencies (Glasberg and Moore, 1990) spanning the range $f \in [f_\mathrm{L}, f_\mathrm{U}]$ and denoted f_c for $c \in [1, N_b]$, where $f_\mathrm{L} = 50$ Hz and $f_\mathrm{U} = 21$ kHz, as recommended by Boren et al. (2015).

For this and other metrics, we further define the *spectral range*, given by

$$\rho_S = \max_c S(f_c) - \min_c S(f_c), \tag{10}$$

and the *spectral deviation*, given by

$$\sigma_S = \sqrt{\frac{1}{N_b}\sum_{c=1}^{N_b}\left(S(f_c) - \overline{S}\right)^2}, \tag{11}$$

where S is some metric (specified in dB, unless stated otherwise) and $\overline{S}$ is its average over all frequency bands. In this case, we define the spectral range and deviation of the ABSE: $\rho_\mathrm{ABSE}, \sigma_\mathrm{ABSE}$, respectively.

ii. Peak and Notch Errors ($E_\mathrm{pk}, E_\mathrm{n}$)

The *peak and notch errors* ($E_\mathrm{pk}, E_\mathrm{n}$) were defined by Boren et al. (2015) and essentially quantify the average peak (or notch) height (depth) in a frequency response over a certain frequency range. First, the difference (in dB) is computed between finely- and coarsely-smoothed versions of the the normalized free-field transfer function $F(f) = F_\mathrm{T}(f)/F_\mathrm{R}(f)$, i.e.,

$$D(f) = 20\log_{10}\left(\frac{\mathcal{S}\left(|F(f)|; 1/48\right)}{\mathcal{S}\left(|F(f)|; 1\right)}\right), \tag{12}$$

where $\mathcal{S}(F;B)$ denotes fractional-octave smoothing[7] applied to the spectrum F with smoothing bandwidth B octaves. The peak- and notch-finding algorithms described by Boren et al. are then applied to find the frequencies $f_1^\uparrow, f_2^\uparrow, \ldots f_{N_\mathrm{pk}}^\uparrow$ of all N_pk spectral peaks and $f_1^\downarrow, f_2^\downarrow, \ldots f_{N_\mathrm{n}}^\downarrow$ of all N_n spectral notches in the range $f \in [f_\mathrm{L}, f_\mathrm{U}]$. The peak and notch errors are then given by (Boren et al., 2015, Eq. (1))

$$E_\mathrm{pk} = \frac{\sum_{j=1}^{N_\mathrm{pk}} D(f_j^\uparrow)}{3\log_2(f_\mathrm{U}/f_\mathrm{L})} \qquad \text{and} \qquad E_\mathrm{n} = \frac{\sum_{j=1}^{N_\mathrm{n}} (-D(f_j^\downarrow))}{3\log_2(f_\mathrm{U}/f_\mathrm{L})}, \tag{13}$$

respectively. Note that, since D is given in dB, the negative sign in the second equation typically ensures that both metrics are positive-valued.

[7] In this work, we used the method described by Tylka et al. (2017).

iii. Central Spectrum (CS)

The *central spectrum* (CS) was defined by Kates (1984) for use as a metric for comparing loudspeaker responses. Consequently, it may be employed using only the free-field transfer functions of the test and reference samples. Specifically, we compute the difference (in dB) between the central spectra for the test and reference samples, given by

$$\mathrm{CS}(f_c) = \mathrm{CS_T}(f_c) - \mathrm{CS_R}(f_c). \tag{14}$$

As done for the ABSE, we define the spectral range and deviation of the CS: $\rho_{\mathrm{CS}}, \sigma_{\mathrm{CS}}$, respectively.

iv. Composite Loudness Level (CLL)

The *composite loudness level* (CLL) spectrum was defined by Pulkki et al. (1999, section 1.1) to give an estimate of perceived timbre. Computing the CLL requires binaural impulse responses, which we compute using Eq. (8). We then compute the difference (in phons) between the CLL spectra for the test and reference samples, given by

$$\mathrm{CLL}(f_c) = \mathrm{CLL_T}(f_c) - \mathrm{CLL_R}(f_c). \tag{15}$$

Again, we define the spectral range and deviation of the CLL: $\rho_{\mathrm{CLL}}, \sigma_{\mathrm{CLL}}$, respectively.

v. Internal Spectrum (IS)

Wittek et al. (2007) adapted the *internal spectrum* (IS) defined by Salomons (1995, chapter 5) in order to define so-called *spectral alterations*. These spectral alterations are computed as the difference (in dB) between the internal spectra for the test and reference samples, given by

$$\mathrm{IS}(f_c) = \mathrm{IS_T}(f_c) - \mathrm{IS_R}(f_c). \tag{16}$$

According to Wittek et al. (2007, section 3.2.5), the IS for each sample is given as the average of the binaural power spectra, i.e.,

$$\mathrm{IS_{T,R}}(f_c) = 10\log_{10}\left(\frac{P^{\mathrm{L}}_{\mathrm{T,R}}(f_c) + P^{\mathrm{R}}_{\mathrm{T,R}}(f_c)}{2}\right). \tag{17}$$

Here, $P^{\mathrm{L,R}}_{\mathrm{T,R}}$ are the binaural power spectra after critical-band filtering, given by (Salomons, 1995, Eq. (5.12))

$$P^{\mathrm{L,R}}_{\mathrm{T,R}}(f_c) = \frac{\int_{-\infty}^{\infty} |C(f;f_c)||B^{\mathrm{L,R}}_{\mathrm{T,R}}(f)|^2 df}{\int_{-\infty}^{\infty} |C(f;f_c)| df}, \tag{18}$$

where $C(f;f_c)$ are Patterson's auditory filters as specified by Salomons (1995, Eq. (5.9)) and $B^{\mathrm{L,R}}_{\mathrm{T,R}}$ are the binaural transfer functions, given by the Fourier transform of the binaural impulse responses from Eq. (8). Again, we define the spectral range and deviation of the IS: $\rho_{\mathrm{IS}}, \sigma_{\mathrm{IS}}$, respectively. Note that ρ_{IS} is precisely equivalent to the A_0-measure defined by Wittek et al. (2007, section 3.2.6), which is based on the A_0-criterion defined by Salomons (1995, section 5.4). Additionally, σ_{IS} is essentially equivalent to the "spectral deviation" described by Wittek et al. (2007, section 3.2.6).

vi. Linear Regression Models

In this work, we used various combinations (listed in Table 1) of the metrics described above to create multiple linear regression models, which predict subjective ratings of coloration, as discussed in Sec. 4.B.

Model Name	Metrics Used
Proposed	$\rho_{\text{ABSE}}, \sigma_{\text{ABSE}}, E_{\text{pk}}, E_{\text{n}}$
Kates (1984)	$\rho_{\text{CS}}, \sigma_{\text{CS}}$
Pulkki et al. (1999)	$\rho_{\text{CLL}}, \sigma_{\text{CLL}}$
Wittek et al. (2007)	$\rho_{\text{IS}}, \sigma_{\text{IS}}$

Table 1: Metrics used for each coloration model.

3. LISTENING EXPERIMENTS

Two listening experiments were conducted in an acoustically-treated listening room, where the listener was seated and given a pair of headphones. Four subjects, all male, ages 25–30 years, participated in the experiments; each subject is an experienced audio engineer or researcher. Prior to the experiments, each subject's HRTFs were measured in an anechoic chamber. These HRTFs were used to render ambisonics to binaural via the ambiX binaural decoder plug-in.[8] In this decoder, we performed a basic (pseudoinverse) ambisonic decoding (Heller et al., 2008, Appendix A.1) for a 36-node Fliege grid and filtered each virtual loudspeaker's signal by the nearest measured HRTF. The headphones (Stax SR-009) were equalized for each subject using a regularized, least-squares equalization filter (Schärer and Lindau, 2009).

The test samples were produced using the ambisonics interpolation techniques of Southern et al. (2009) and Tylka and Choueiri (2016), employed for microphone spacings of 10, 30, or 50 cm. The test samples were rendered from 4$^{\text{th}}$-order HOA room impulse responses for microphone positions distributed on both sides of the listening position, as illustrated by the empty circles in Fig. 1. Also included in each test were reference samples, measured at the listening position and for which no interpolation was performed.

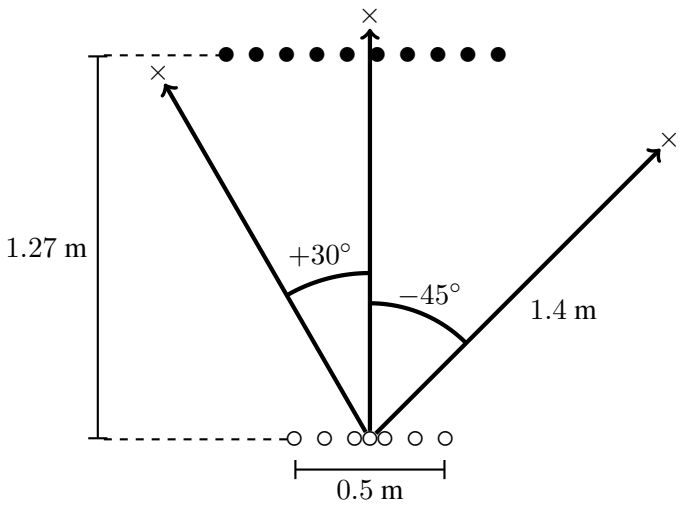


Figure 1: Diagram of microphone and source positions used in the listening tests. The empty circles indicate the microphone positions; the filled circles indicate the source positions used in the localization test; and the crosses indicate the source positions used in the coloration test.

[8] Available here: `http://www.matthiaskronlachner.com/?p=2015`

A. LOCALIZATION TEST

To measure perceived source localization, we conducted a virtual source localization test, in which the listener was seated 1.27 m in front of a horizontal linear array of 30 transducers spaced 5 cm apart (note that the array served only as a visual reference to promote sound externalization). An infrared head-tracking device (NaturalPoint TrackIR) was used to maintain a stable sound field as the subject's head rotated. The test consisted of 1 training round followed by 5 rounds of testing, with optional short breaks in between each round for the subject to stand up and take off the headphones. In each round, the subject was presented with 14 randomly-selected samples (2 references and 12 test samples in each round), all of which were a short ($\sim$ 2 second) clip of male English speech. The intended source directions produced in the test corresponded to 10 of the 30 transducers, spanning approximately $\pm 20^\circ$ azimuth on the horizontal plane, as illustrated by the filled circles in Fig. 1. The subject was asked to identify the direction from which the sound appeared to originate, and subsequently face it. The subject's head-angle (obtained from the head-tracking device) was then captured as the perceived direction of the source. The subject was able to repeat each sample any number of times until confident about the location of the source.

B. COLORATION TEST

To collect subjective ratings of coloration, we conducted an ITU-R BS.1534-3 MUSHRA test, administered in the same listening room and with the same headphones, but without head-tracking. The test consisted of 1 training round followed by 3 rounds of testing, with optional short breaks between rounds. In each round, the subject was presented with 9 "test samples" (actually 6 test samples, 2 anchors, and a hidden reference) and a labeled reference sample, all of which were a short ($\sim$ 3 second) clip of pink noise. The low anchor was the standard 3.5 kHz low-pass-filtered version of the reference; the second anchor was a high-shelf-filtered version of the reference, with $+6$ dB of gain applied above 7 kHz. The samples were randomly-ordered, but all samples in each round were from a single intended source direction. The intended source directions produced in the test corresponded to -45°, 0°, 30° azimuth on the horizontal plane, as illustrated by the crosses in Fig. 1. The subject was asked to judge (and rate on a scale from 0–100) the extent to which each test sample *differs*, in terms of the tonal coloration only, from the reference. As is standard in a MUSHRA test, a rating of 100 indicates that the sample is *indistinguishable* from the reference, while any rating less than 100 indicates that the sample differs from the reference. All responses for each round and each subject were mapped (if necessary) such that the low anchor obtained a rating of 0. In the present dataset, all subjects correctly identified the hidden reference and rated it 100. The subject was able to repeat each sample (and the reference) any number of times until satisfied with the ratings for that round.

4. RESULTS

Using the data collected in the listening experiments, we determined optimal localization model parameters, as described in Sec. 4.A, and constructed the coloration models, as described in Sec. 4.B.

A. COMPARISON OF LOCALIZATION MODELS

Using the measured localization directions from the localization experiment (described in Sec. 3.A), we first determined optimal values for each of the 4 parameters discussed in Sec. 2.A.i. The optimization consisted of minimizing the squared-residuals between the predicted and measured localization azimuths. The resulting parameter values are listed in Table 2. From these optimal values, we see that, generally, the low-frequency (velocity) vector requires a "coarser" set of input data, as the spatial resolution (related to Q) is much lower. Conceptually, this agrees with the notion that low-frequency sounds are not very directional, so a low-spatial-resolution representation of such information should be adequate. Similarly, the wavelet

Parameter	Velocity Vector	Energy Vector	Description
Q	9	36	Number of plane-waves
α	0.95	0.7	Stationary signal weight
τ (ms)	2	1	Minimum wavelet half-length
γ (dB)	-8	-30	Wavelet detection threshold

Table 2: Optimal parameters for each localization vector. More detailed descriptions of each parameter are given in Sec. 2.A.i.

Model Name	Residuals	Correlation	$\bar{\epsilon}$
Proposed	706.5	0.77	3.67°
Dietz et al. (2011)	1009.2	0.82	4.34°

Table 3: Squared residuals, Pearson correlation coefficients, and mean absolute prediction errors ($\bar{\epsilon}$) for each localization model. The squared residuals are normalized by the variance seen in the measured data for each sample.

lengths (set by τ) are longer for the velocity vector than for the energy vector, which is likely a result of low-frequency information requiring longer time-scales in order to be adequately represented.

In Fig. 2, the measured localization directions are plotted against the predictions of each model. The mean absolute prediction error is given by

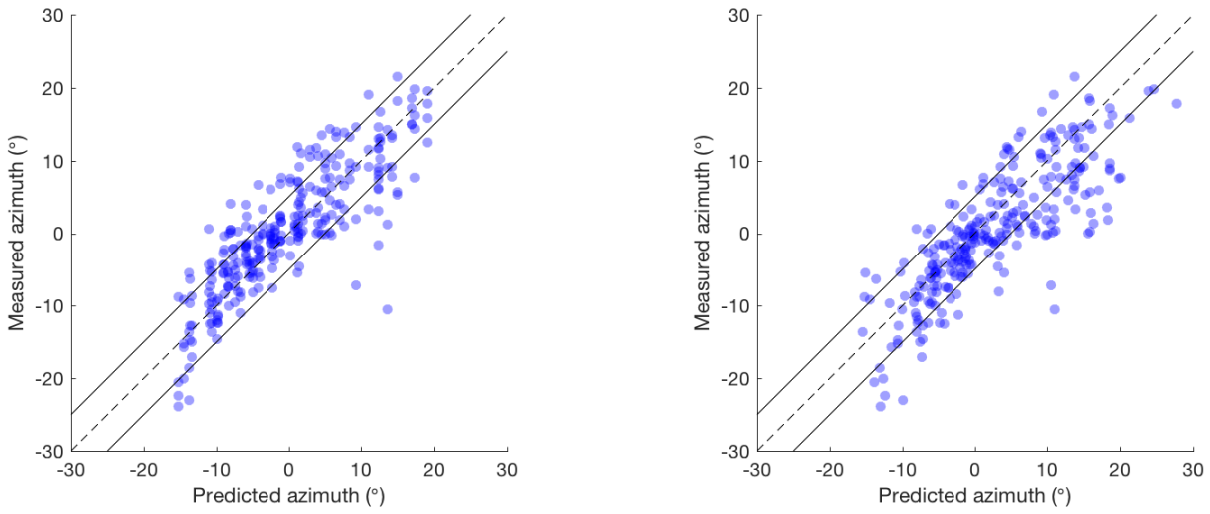
$$\bar{\epsilon} = \frac{1}{R}\sum_{r=1}^{R} |\theta_r - \theta_p|, \tag{19}$$

where R is the total number of responses, θ_r is the measured azimuth for response r, and θ_p is the predicted azimuth. These errors, as well as the squared residuals and Pearson correlation coefficients for the data, are given in Table 3. From these values, we see that while the proposed model seems to fit the data better compared to the binaural model, as the former achieves both a lower squared-residual value as well as a smaller mean absolute prediction error, the latter achieves a higher correlation with the data. This may be explained (in part) by the binaural model's ability to take into account subject-dependent variations, since the predictions are made on a per-subject basis (see Sec. 2.A.ii), as well as any effects of the binaural rendering approach used. The proposed model, however, is unaware of the binaural rendering approach used and can only make a single prediction per sample, meaning that any subject-dependent variation in the data cannot be captured.

For both models we observe two outlying data points at approximately $(+10°, -10°)$, for which the predicted directions are to the left $(+)$, while the subject localized the sound to the right $(-)$. Although not shown here, the same data points appear again as outliers when plotted against *intended* source direction, suggesting these outliers could well be due to an error on the part of the subject.

B. COMPARISON OF COLORATION MODELS

Using the MUSHRA ratings collected from the coloration listening test (described in Sec. 3.B), we performed linear regressions for each model discussed in Sec. 2.B.vi. We first converted the MUSHRA ratings to "coloration scores", given by $C = 100 - M$, where M are the MUSHRA ratings. Through this transformation, a reference sample will always have a coloration score of zero, while the low-pass anchor will have a coloration score of 100. We then computed linear regressions between the values of the metrics



(a) Proposed model ***(b) Dietz et al. (2011) model***

Figure 2: Scatter plots of measured versus predicted source directions. The filled blue circles indicate the data and model values; the dashed black lines indicate ideal model results (i.e., $\theta_r = \theta_p$); the solid black lines indicate discrepancies of 5° (i.e., $\theta_r = \theta_p \pm 5^\circ$).

in each of the four models and the coloration scores. Note that, for the binaural models (Pulkki et al., 1999; Wittek et al., 2007), we computed each metric on a per-subject basis, i.e., using each subject's individualized HRTFs to render to binaural and subsequently computing per-subject values of each metric.

In building the coloration models, an analysis of the statistical significance of each model parameter revealed that, for the proposed model, neither σ_{ABSE} nor E_{pk} provided a significant improvement to the model. This is likely because, for all of our test samples, both σ_{ABSE} and E_{pk} were strongly correlated with ρ_{ABSE}. Consequently, our proposed model uses only ρ_{ABSE} and E_{n}. Additionally, a "y-intercept" (offset) term was considered for each model, but was found to be statistically insignificant in all cases.[9] The final formulae and corresponding Pearson correlation coefficients between the measured and predicted coloration scores are tabulated in Table 4. These correlation coefficients suggest that the proposed model is best able to predict the measured coloration scores.

The proposed model is also the only model to have both coefficients positive, even though all of the metrics listed in Sec. 2.B produce positive values for non-flat spectra. This indicates that both metrics in the proposed model directly contribute to perceived coloration. We also note the similarity between the two binaural models (Pulkki et al., 1999; Wittek et al., 2007), in both the coefficients and performance of the model. This may be explained by both models capturing essentially the same information through combining the binaural spectra.

Additionally, in Fig. 3, the measured coloration scores are plotted against the predicted scores for each model. From these plots, we see that the proposed model produces the most compact (towards the $y = x$ line) distribution of the data. We also note that the model of Kates (1984) is the only model that consistently under-predicts the low-anchor scores. This suggests that this model did not have sufficient degrees of freedom (since the two metrics were strongly correlated with one another) to capture the end points of the data.

[9]This is likely because all of the reference samples, by definition, obtain zero coloration scores and zero values for each metric.

Model Name	Formula	Correlation
Proposed	$C = 2.88\rho_{\text{ABSE}} + 1.74E_{\text{n}}$	0.84
Kates (1984)	$C = -3.20\rho_{\text{CS}} + 54.16\sigma_{\text{CS}}$	0.72
Pulkki et al. (1999)	$C = 10.79\rho_{\text{CLL}} - 19.75\sigma_{\text{CLL}}$	0.77
Wittek et al. (2007)	$C = 10.36\rho_{\text{IS}} - 19.62\sigma_{\text{IS}}$	0.77

Table 4: Formulae and Pearson correlation coefficients for each coloration model. All correlation p-values were less than 10^{-17}.

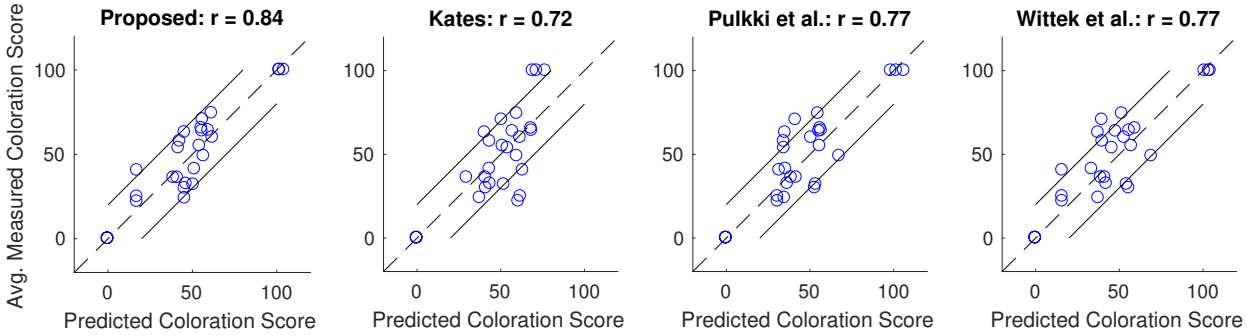


Figure 3: Scatter plots of average measured versus predicted coloration scores for each model. The blue circles indicate the data and model values; the dashed black lines indicate ideal model results (i.e., $y = x$); the solid black lines indicate discrepancies of 20 points (i.e., $y = x \pm 20$). The predicted coloration scores for the binaural models (Pulkki et al., 1999; Wittek et al., 2007) are averaged across listeners. Correlation coefficients for each set of data are given at the top of each plot.

5. CONCLUSIONS

In this work, we developed models for perceived source localization and spectral coloration. We empirically determined parameters of these models through comparison with results of subjective listening experiments. One advantage of these models, compared to existing ones, is that they do not require rendering ambisonics to binaural. This allows the models to be used to directly evaluate navigational techniques for higher-order ambisonics, without introducing extraneous factors such as the choice of ambisonics-to-binaural rendering approach or of HRTF.

The localization model (described in Sec. 2.A.i) extends a recently-developed precedence-effect-based energy vector model (Stitt et al., 2016) in order to predict perceived source localization directly (i.e., without rendering to binaural) from the ambisonics impulse responses. To determine parameters of the model, we conducted a virtual localization test (described in Sec. 3.A) with individualized binaural rendering over head-tracked and equalized headphones. The results of the localization test are in good agreement with the predictions of the localization model (see Sec. 4.A), achieving a mean absolute prediction error of 3.67°. Furthermore, the proposed model performs comparably to, if not better than, the binaural localization model of Dietz et al. (2011) (described in Sec. 2.A.ii), in terms of their agreement with the data.

The coloration model uses only the omnidirectional ambisonics impulse responses (i.e., the free-field transfer functions) and predicts a perceived "coloration score" from a linear combination of two metrics (defined in Sec. 2.B): the range of the auditory band spectral error (ρ_{ABSE}) and the notch errors (E_{n}). To construct this model, we conducted a MUSHRA (ITU-R BS.1534-3) test (described in Sec. 3.B) and per-

formed a linear regression of the metrics with the collected subjective ratings of coloration. We compared the proposed model to several alternative models and found it to achieve the highest correlation to the measured data (see Sec. 4.B).

It should be noted, however, that predicting these "coloration scores" is a somewhat artificial task, as the scale (0–100) is arbitrary, and there is no reason to think that the perceived coloration should be strictly linearly related to any of the metrics used. Nevertheless, a more general result of this analysis is that the metrics used in the proposed model (ρ_{ABSE} and E_{n}) are dominant factors in the perception of coloration. Thus, each of these metrics may serve as a useful measure of perceptible spectral coloration, as a large value for either metric would almost certainly yield perceptible coloration.

A. FUTURE WORK

To further validate and refine the proposed models, additional listening experiments should be conducted with more subjects, additional source positions, and varied stimuli. For example, the localization model has only been validated for frontal ($\pm 20^\circ$ azimuth) sources and a speech signal. Furthermore, the stimulus-dependent stationary signal weight (α) is typically determined empirically by fitting model predictions to experimental data (see Stitt et al., 2016, 2017). Consequently, future work should seek a (possibly empirical) model for α such that it can be determined *a priori* for a given stimulus.[10]

Similarly, the coloration model should be validated against other navigational techniques, as the spectral colorations induced by the techniques examined here may or may not be comparable to those induced by alternative techniques. Finally, alternative binaural rendering techniques should be employed in future listening experiments in order to verify the desired insensitivity of the models to choice of rendering approach.

Despite their need for further validation, the models presented here appear promising for use in comparing navigational techniques. Consequently, a comprehensive comparison of existing navigational techniques should be conducted using these models in order to quantify the penalties incurred by each technique, and ultimately determine limits of usability for each technique (e.g., maximum translation distance with $\leq 5^\circ$ source localization error).

ACKNOWLEDGMENTS

The HOA room impulse responses were recorded using the Eigenmike by mh Acoustics.[11] The authors wish to thank P. Stitt for providing the MATLAB code for the precedence-effect-based energy vector model.[12]

REFERENCES

S. Berge and N. Barrett. High angular resolution planewave expansion. In *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, May 2010. URL `http://ambisonics10.ircam.fr/drupal/index5d87.html?q=proceedings/o15`.

B. Boren, M. Geronazzo, F. Brinkmann, and E. Choueiri. Coloration metrics for headphone equalization. In *Proceedings of the 21st International Conference on Auditory Display*, pages 29–34, July 2015. URL `http://iem.kug.ac.at/fileadmin/media/institut-17/icad15/proceedings150707_2.pdf`.

[10] Stitt et al. (2016) suggest a simple model for α based on interaural time differences of leading and lagging signals, but no direct relationship between the stimulus and α is given.

[11] See: `https://www.mhacoustics.com/products#eigenmike1`

[12] Available here: `https://spatialaudio.xyz/matlab-code/`

M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, 2011. ISSN 0167-6393. doi: 10.1016/j.specom.2010.05.006. URL http://www.sciencedirect.com/science/article/pii/S016763931000097X. Perceptual and Statistical Audition.

R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis. High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues. In *Audio Engineering Society Convention 119*, October 2005. URL http://www.aes.org/e-lib/browse.cfm?elib=13369.

J. Fliege and U. Maier. The distribution of points on the sphere and corresponding cubature formulae. *IMA Journal of Numerical Analysis*, 19(2):317–334, 1999. doi: 10.1093/imanum/19.2.317.

M. A. Gerzon. General metatheory of auditory localisation. In *Audio Engineering Society Convention 92*, March 1992. URL http://www.aes.org/e-lib/browse.cfm?elib=6827.

B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103–138, 1990.

N. A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier Science, 2005.

N. Hahn and S. Spors. Physical properties of modal beamforming in the context of data-based sound reproduction. In *Audio Engineering Society Convention 139*, October 2015. URL http://www.aes.org/e-lib/browse.cfm?elib=18024.

A. Heller, R. Lee, and E. Benjamin. Is my decoder ambisonic? In *Audio Engineering Society Convention 125*, October 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14705.

ITU-R BS.1534-3. Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems, 2015.

J. M. Kates. A perceptual criterion for loudspeaker evaluation. *The Journal of the Audio Engineering Society*, 32(12):938–945, 1984. URL http://www.aes.org/e-lib/browse.cfm?elib=4469.

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*, 80(6):1608–1622, 1986. doi: 10.1121/1.394325.

C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi. ambiX - A suggested ambisonics format. In *Proceedings of the 3rd Ambisonics Symposium*, June 2011. URL http://ambisonics.iem.at/proceedings-of-the-ambisonics-symposium-2011/ambix-a-suggested-ambisonics-format.

V. Pulkki, M. Karjalainen, and J. Huopaniemi. Analyzing virtual sound source attributes using a binaural auditory model. *The Journal of the Audio Engineering Society*, 47(4):203–217, 1999. URL http://www.aes.org/e-lib/browse.cfm?elib=12110.

A. M. Salomons. *Coloration and Binaural Decoloration of Sound due to Reflections*. PhD thesis, Delft University of Technology, 1995.

Z. Schärer and A. Lindau. Evaluation of equalization methods for binaural signals. In *Audio Engineering Society Convention 126*, May 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=14917.

F. Schultz and S. Spors. Data-based binaural synthesis including rotational and translatory head-movements. In *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control - Engineering and Perception*, September 2013. URL http://www.aes.org/e-lib/browse.cfm?elib=16894.

P. L. Søndergaard and P. Majdak. The auditory modeling toolbox. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 33–56. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-37762-4. doi: 10.1007/978-3-642-37762-4_2.

A. Southern, J. Wells, and D. Murphy. Rendering walk-through auralisations using wave-based acoustical models. In *Signal Processing Conference, 2009 17th European*, pages 715–719, August 2009.

P. Stitt, S. Bertet, and M. van Walstijn. Extended energy vector prediction of ambisonically reproduced image direction at off-center listening positions. *The Journal of the Audio Engineering Society*, 64(5): 299–310, 2016. URL http://www.aes.org/e-lib/browse.cfm?elib=18135.

P. Stitt, S. Bertet, and M. van Walstijn. Off-center listening with third-order ambisonics: Dependence of perceived source direction on signal type. *The Journal of the Audio Engineering Society*, 65(3):188–197, 2017. URL http://www.aes.org/e-lib/browse.cfm?elib=18554.

J. G. Tylka and E. Y. Choueiri. Comparison of techniques for binaural navigation of higher-order ambisonic soundfields. In *Audio Engineering Society Convention 139*, October 2015. URL http://www.aes.org/e-lib/browse.cfm?elib=17977.

J. G. Tylka and E. Y. Choueiri. Soundfield navigation using an array of higher-order ambisonics microphones. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, September 2016. URL http://www.aes.org/e-lib/browse.cfm?elib=18502.

J. G. Tylka, B. B. Boren, and E. Y. Choueiri. A generalized method for fractional-octave smoothing of transfer functions that preserves log-frequency symmetry. *The Journal of the Audio Engineering Society*, 65 (3):239–245, 2017. doi: 10.17743/jaes.2016.0053. URL http://www.aes.org/e-lib/browse.cfm?elib=18558.

H. Wierstorf, A. Raake, and S. Spors. Binaural assessment of multichannel reproduction. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 255–278. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-37762-4. doi: 10.1007/978-3-642-37762-4_10.

F. Winter, F. Schultz, and S. Spors. Localization properties of data-based binaural synthesis including translatory head-movements. In *Forum Acusticum*, September 2014.

H. Wittek, F. Rumsey, and G. Theile. On the sound color properties of wavefield synthesis and stereo. In *Audio Engineering Society Convention 123*, October 2007. URL http://www.aes.org/e-lib/browse.cfm?elib=14225.