Audio Engineering Society

Conference Paper

Presented at the International Conference on Audio for Virtual and Augmented Reality, 2020 August 17–19, Redmond, WA, USA

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Quantifying HRTF Spectral Magnitude Precision in Spatial Computing Applications

Steven Crawford[1], Rémi Audfray[2], and Jean-Marc Jot[2]

[1]*University of Rochester* - [2]*Magic Leap, Inc.*

Correspondence should be addressed to Steven Crawford (s.crawford@rochester.edu)

ABSTRACT

In this paper, an algorithmic approach towards computing quantifiable metrics regarding HRTF spectral magnitude synthesis performance of virtual sound systems, such as those used in VR/AR/MR environments, is presented. Utilizing regularized regression in parallel with a statistical information theory technique, the system provides a detailed analysis of a virtual spatializer's spectral magnitude rendering accuracy at a given point in space. Applying the proposed system to the final signal processing stage of a spatial audio rendering pipeline enables the engineer to establish critical performance quantities for benchmarking future modifications to the rendering channel against. The proposed system demonstrates an important step towards standardizing and automating virtual audio system evaluation and may ultimately act as a participant substitute during critical listening tasks.

1 Introduction

Conventional approaches towards virtual audio rendering system development have partially relied on perceptual evaluation for determining and comparing the efficacy of various rendering schemes and reproduction methods. Notwithstanding the prerequisite necessity for perceptual critical listening evaluation in order to ultimately qualify audio rendering performance, the importance of developing computational tools for engineers attempting to quantify system performance is likewise essential. Perceptual listening evaluations can be costly in terms of both time and resources and quantitative signal-based analysis has the potential to mitigate some of the dependence on human listeners. Audio quality standard practice recommendations regarding perceptual audio encoding/decoding artifacts, speech intelligibility, and/or compression/transmission distortions are well established [1, 2, 3, 4, 5, 6]. However, in regards to the developing field of VR/AR/MR audio systems, recent attempts towards computational quantification are somewhat sparser [7, 8, 9]. This presents a compelling impetus and catalyst for the investigation and development of an assortment of quantitative tools.

Ideally, we desire an objective metric computed using empirical features from the audio signals that can estimate the perceived audio quality reported by a trained listener. These types of systems may ultimately be able to act as reliable artificial listeners in a variety of perceptual psychoacoustic testing, and as such serve as tools aiding in the development and refinement of spatial audio rendering systems. In this paper, we present an approach at quantifying the

synthesis accuracy of head related transfer function (HRTF) spectral magnitude reproduction within a virtual audio rendering system (termed 'spatializer'). In Section 2, we provide an overview of some general requirements for such a quantitative audio analysis tool, briefly describe human spatial audition as it relates to the standard signal flow in a typical spatializer setup, and present a description of the proposed system architecture. Section 3 presents applied examples of the proposed system using a generic cube-based virtual loudspeaker array, and Section 4 concludes with a discussion and possible applications of the system in the future.

2 Background and System Overview

Conventionally speaking, we are interested in a 'full-reference' or 'comparison-based' computational measurement method in which we compare an unaltered test signal with a version of the signal that has undergone arbitrary processing by some system [3]. Fundamentally, this amounts to combining some type of signal transform modeling human hearing and/or psychoacoustic perception, along with a distance coefficient computed between the two signals.

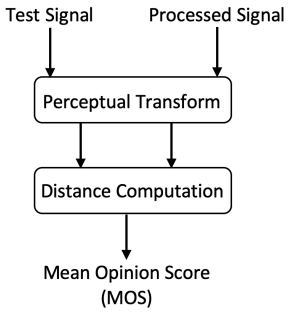

Figure 1. First-principles of a generic full-reference measurement system.

Figure 1 displays a high-level signal flow diagram for a generic full-reference computational measurement system. Both the 'Perceptual Transform' and 'Distance Computation' processing Blocks are intended to be agnostic towards the model chosen by the engineer. The output from the Distance Computation Block is a metric (mean opinion score - MOS) that ideally represents the average "audio quality" that would be reported by a group of trained listeners [10].

2.1 Spatial Auditory Perception

Sound source localization depends on 3 primary cues; interaural time differences (ITD), interaural level differences (ILD), and monaural spectral cues. When a sound source is located to the right of a listener, the sound wave emitted by the source arrives at the right ear (ipsilateral – same side) first, then diffracts and travels the extra distance around the head to the left ear (contralateral – opposite side). The time it takes the wave to travel the extra distance around the head is called the ITD, which for low frequencies (around 1.5kHz or less), is the most critical cue used in horizontal localization [11]. This same scenario also produces ILDs via acoustic-shadowing (i.e. low pass filtering) caused by the head and torso of the listener. Interaural level differences are functions of spatial orientation and assist in azimuthal localization at frequencies above ~1.5kHz. When a sound source is on the median plane, ITD and ILD are zero and when a sound source is laterally located, they are maximised.

The predominant cues used to determine the elevation of a sound source are spectral in nature [12, 13]. The anatomy of the outer ear produces idiosyncratic spatial filtering with strong dependencies on frequency and sound source angle of incidence [14]. The raised and folded areas of the pinna act as chambers in which shorter wavelength frequencies undergo directionally dependent interference patterns, producing spectral notches which increase in frequency with elevation. Over time, an individual's auditory system learns to interpret the spectral notches produced by their specific anatomy and this aids greatly in vertical sound source localization [15]. The total impact that an individual's shoulders, neck, head, and external ears have on both monaural spectral and binaural difference localization cues is captured in their individual head-related impulse response (HRIR). The HRIR is dependent on both azimuth and elevation as well as source distance.

Transforming the HRIR into the frequency domain produces the complex valued head-related transfer

function (HRTF). The HRTF models the spatial filtering a propagating sound wave undergoes along its path from source-to-ear. The absolute value of the HRTF defines the frequency magnitude response, the argument designates the phase response, and the ratio of contralateral to ipsilateral responses defines the interaural transfer function (ITF). Upon decomposing the HRTF into a minimum-phase system cascaded with a delaying all-pass element, the resulting interaural excess phase difference (IEPD) can be used as a highly accurate estimator of ITD across frequency [16, 17].

2.1.1 Spatializer Fundamentals

The overarching and underlying goal of any virtual audio spatializer system is to faithfully reproduce the same acoustic pressure and velocity at the ears of a listener that would have organically occurred in the actual listening environment being simulated. Typical real world audio systems rely on loudspeaker arrays and amplitude panning schemes to collectively reproduce interaural difference and spectral cues at a centrally located listening position, and thus creating the virtual auditory image. A virtual spatializer simulates a loudspeaker array using HRIR filters corresponding with the intended locations of the simulated loudspeakers.

By filtering a monophonic sound source with a summed and weighted linear combination of HRIRs, a 3D virtual auditory image can be binaurally rendered (or 'spatialized') over headphones [19]. Just as in real world loudspeaker array and panning scheme applications, the virtual array produces auditory images using discrete amplitude panning techniques to smoothly interpolate between loudspeaker positions. There are several factors determining the overall perceived quality of the spatialized simulation (e.g. HRIR selection, measurement method, and HRTF equalization techniques, headphone equalization methods, head-tracking, multimodal cues, simulated reflections and reverberation, room models, etc.), and a full discussion is outside the scope of this paper; for a comprehensive review on spatialization technologies and techniques, see [16-19, 21, 22].

2.2 System Architecture Overview

The system described here makes use of two different 'Distance Computation' Blocks (Fig. 1) and uses dB-weighted HRTF filtering to account for the 'Perceptual Transform' Block. The first distance computation instance uses regularized regression (an elastic-net), while the second instance uses a probabilistic information theory technique, the Jensen-Shannon distance. The high-level algorithmic approach of the system is founded upon a set of test signal functions serving as a "ground-truth" against which an arbitrary processed signal can be compared. The test set is composed of a database of HRIRs, individualized or generic, measured or synthesized over a discrete set of spatial points on a sphere surrounding the listener. The measurement resolution of the chosen test set (i.e. HRIR database) will determine the resolution of the returned MOSs.

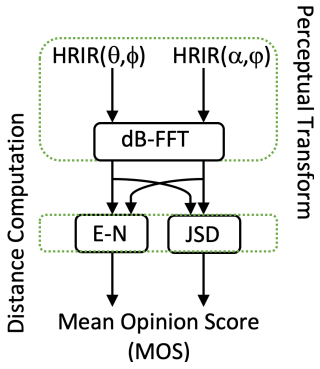

Figure 2. System architecture signal flow diagram.

The processed signal that the system operates on is a head related impulse response produced at an arbitrary location in space – HRIR(α,φ). This HRIR may be completely unrelated to the test set, or it may be a modified and/or processed version of a function, or linear combination of functions from the test set. Both the processed HRIR(α,φ) and the test set HRIR(θ,ϕ) are transformed into the frequency domain by Fourier transform, placed on a log (i.e. magnitude-dB) scale, and then compared to one another using the elastic-net and Jensen-Shannon distances. This method is repeated between the processed HRIR (with constant (α,φ) values) and every (θ,ϕ) combination available within the HRIR test signal set; each block producing a MOS every

iteration (Fig. 2). Rendering deficiencies in either the time or frequency domains will produce undesirable effects to the spatialized sound source and both domains must be considered to form a complete metric set. However, the current analysis assumes accurate spatializer reproduction of ITD cues and thus focuses on the rendering accuracy of spectral magnitude cues. This assumption may not hold true in all circumstances and therefore, future investigation into ITD evaluation and analysis of the IEPD quantity is warranted [25].

2.2.1 Elastic-Net Regression

Elastic-net regression incorporates penalties from both the L1-norm and L2-norm into a single cost function according to Eq. 1. In our case, $\boldsymbol{y}$ is the processed signal, $\boldsymbol{X}$ is a matrix where each column is an HRIR from the test set, and the variables δ and λ are shrinkage parameters on the L2 and L1-norms respectively.

$$\hat{\beta} = argmin_{\beta}||\boldsymbol{y} - \boldsymbol{X}\beta||^2 + \delta||\beta||^2 + \lambda||\beta||_1 \quad (1)$$

The L1-norm produces sparsity within the model, while the L2-norm promotes grouping by removing the limitation on the number of variables selected in the regression process and stabilizing the L1-path [23]. In this way, the elastic-net performs a type of grouped variable selection not possible with either the L1-norm or L2-norm alone [24]. The elastic-net enables the system to identify similarities between a processed HRIR and significant groupings of test set HRIRs within a generally sparse model. Similarly, the elastic-net allows us to remove correlated groupings of predictor variables from the test signal set that are not statistically significant to the processed input response signal. The elastic-net returns a vector of coefficients, $\hat{\beta}$, computed between the processed response signal HRIR(α,φ), and every member of the test HRIR set. The returned $\hat{\beta}$ coefficients represent the load on each predictor from the test HRIR set required to linearly recombine and reproduce, as closely as possible, the input response vector.

2.2.2 Jensen-Shannon Distance

The Jensen-Shannon distance (JSD), a finite bounded and symmetrized version of the Kullback-Leibler (KL) divergence, is related to the mutual information between two distributions [26]. JSD is a measurement of similarity between two distributions and serves a distance metric within the proposed quantitative system. In this sense, the JSD considers the processed and test signal inputs as probability distributions from given random variables. The JSD computes the distance between the processed HRIR, P, and every member of the test set, T, according to Eq. 2. The lower the value of the computed JSD, the more statistically similar the two distributions are, with a JSD value of zero meaning the distributions are identical.

$$JSD(P||T) = \sqrt{\frac{1}{2}[KL(P||\frac{P+T}{2}) + KL(T||\frac{P+T}{2})]} \quad (2)$$

$$KL(P||T) = \sum P(x) \log(\frac{P(x)}{T(x)}) \quad (3)$$

3 System Application

3.1 Distance Computation I: E-N

The HRIR test signal set is founded upon measurements taken on the KU100 binaural head from the Sadie database (i.e. v.1 – subject 002). There is psychoacoustic evidence in the literature supporting the notion that this generic HRIR set performs well over a broad variety of listeners [27]. In the following sections, VBAP is used as the amplitude panning scheme to create auditory images using a virtual loudspeaker array (VSA) based on a modified cube, as depicted in Fig. 3 [28, 29]. In Fig. 3, the x-axis displays azimuth locations in 5° increments, while the y-axis displays polar locations (elevations) in 10° increments.

Before running the system, the input test signals are assumed to be standardized/normalized to unit-length and zero-mean, and the processed signal is assumed to have been centered to zero-mean. The system may be calibrated to a virtual loudspeaker position by first tuning the value of δ (Eq. 1) to produce the desired response, while using cross-validation to select the optimal corresponding value of λ (Eq. 1) which minimizes the mean squared error. After computing Eq. 1 between the processed

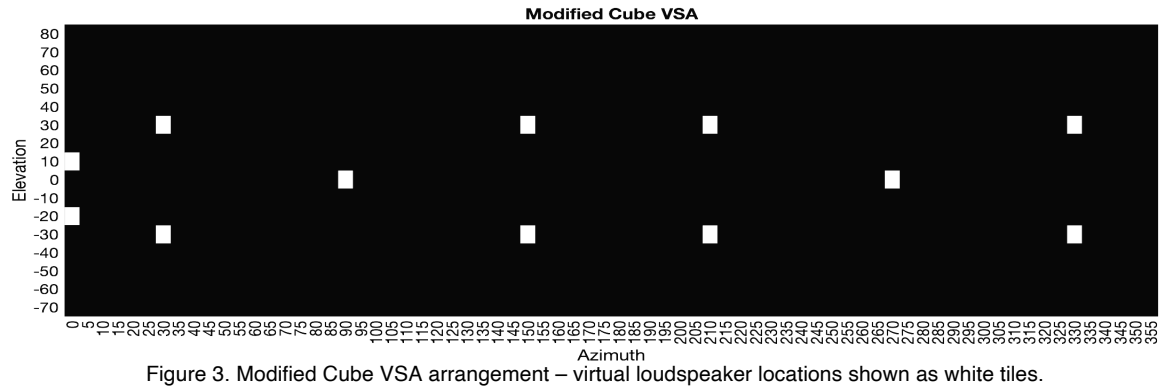

Figure 3. Modified Cube VSA arrangement – virtual loudspeaker locations shown as white tiles.

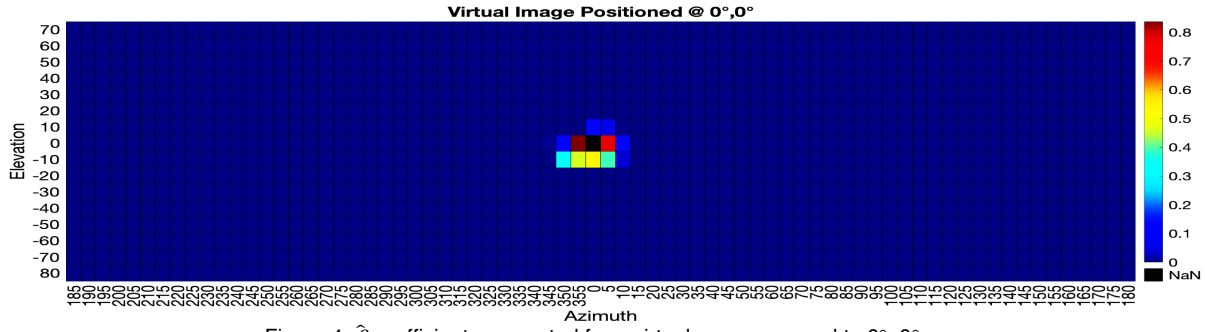

Figure 4. $\hat{\beta}$ coefficients computed for a virtual source panned to 0°, 0°.

response signal HRIR(α,φ) and every member of the test signal set HRIR(θ,ϕ) (where θ spans 0° to 355° and ϕ spans -70° to 80°), the returned $\hat{\beta}$ coefficient weights are indexed by θ and ϕ. The quantities of interest here are the centroid and standard deviation of the distribution of $\hat{\beta}$ coefficients, along with the angular distance from the computed centroid to the intended spatial rendering position. The centroid location is an estimation of the localization point for the virtual auditory image. However, prior to computing the coefficient centroid, it is necessary to shift the entire distribution to the center of the image.

Figure 4 shows the centered (as evidenced by comparing the azimuth and elevation axes of Fig. 3 and Fig. 4) distribution of coefficients for a virtual source panned directly in front of the listening position (i.e. 0°, 0°), and the computed centroid is displayed as the black tile (NaN). The first MOS value produced by the E-N Block, referred to as MOS-1, regards the angular distance from the computed centroid to the intended spatialization location as an estimation of localization accuracy. The second MOS value from the E-N Block, MOS-2, regards the standard deviation of the returned $\hat{\beta}$ coefficients as an image quality estimator, with a larger deviation indicative of a more diffuse virtual image and a smaller deviation indicative of a more compact image.

In Fig. 5, the final output from the E-N Block is shown for a virtual source panned to 0°, 0°, and the computed values for MOS-1 and MOS-2 are 0° and 14.09°, respectively. An MOS-1 value of 0° is ideal and means that the computed centroid of the distribution is equivalent to the intended spatial rendering location. The MOS-2 value, being the standard deviation from the mean, is used to define the radius of a circle centered around the computed centroid of the distribution. Figure 5 visually demonstrates how MOS-2 can be used to estimate the compactness or diffuseness of a rendered virtual auditory image. Figure 6 shows the output of the E-N Block for a virtual image amplitude panned to 120°, 10°.

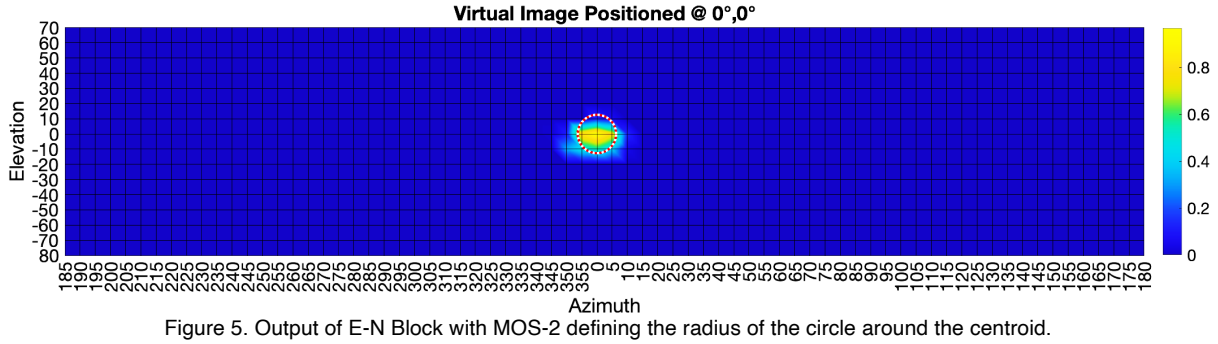

Figure 5. Output of E-N Block with MOS-2 defining the radius of the circle around the centroid.

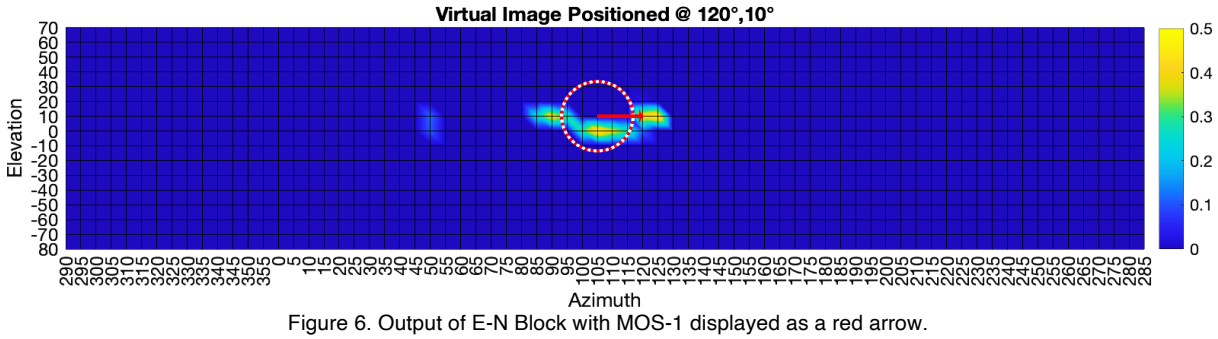

Figure 6. Output of E-N Block with MOS-1 displayed as a red arrow.

In Fig. 6, MOS-1 is graphically represented as a red arrow whose tail originates at the computed centroid (i.e. the estimated spatialization location) and whose tip terminates at the intended spatialization location. In this instance, the computed values for MOS-1 and MOS-2 are 15° and 30.52°, respectively. This means that the virtual image created at this position, using VBAP to pan on the modified cube VSA, is about twice as diffuse as the image created on the front-center line. Additionally, the virtual auditory image created is estimated to be perceived at an azimuth of 105° and an elevation of 10°, which is a 15° difference from the intended spatialization location.

3.2 Distance Computation II: JSD

The JSD 'distance computation' Block also produces two MOS values representative of localization estimation (MOS-3) and auditory image compactness or diffuseness (MOS-4). However, the signal processing flow producing the MOS values in the JSD Block is slightly modified from the MOS computation procedure in the E-N Block. The JSD Block may be calibrated by using a signal from the test set HRIR(θ,ϕ) as both the processed and the test signal. First, the JSD is computed between the processed signal - HRIR(α,φ), and every member of the test signal set HRIR(θ,ϕ), according to Eq. 2. Then, each computed JSD value is subtracted from 1, and plotted according to azimuth and elevation. Figure 7 displays the first-stage output of the JSD Block for a calibration phase using HRIR(θ,ϕ) = HRIR(α,φ) = (90°, 0°). In Fig. 7, a value of 1 indicates identical distributions and a value of 0 indicates zero statistical correlation.

After computing the JSD for each comparison, the angular distance map (ADM), defined as the angular distance between the intended spatialization location and every other location, is computed and shown in Fig. 8. In Fig. 8, the values of angular distances are normalized in between 0-1, and a value of 0 indicates no spatial separation, while a value of 1 indicates an angular separation of 180°. MOS-3, the localization estimator, is taken as the value at the index of the ADM corresponding to the primary return index (i.e. the highest coefficient value) from the JSD. Finally, the computed JSD coefficient values are each multiplied by the corresponding spatial location values taken from the ADM, and

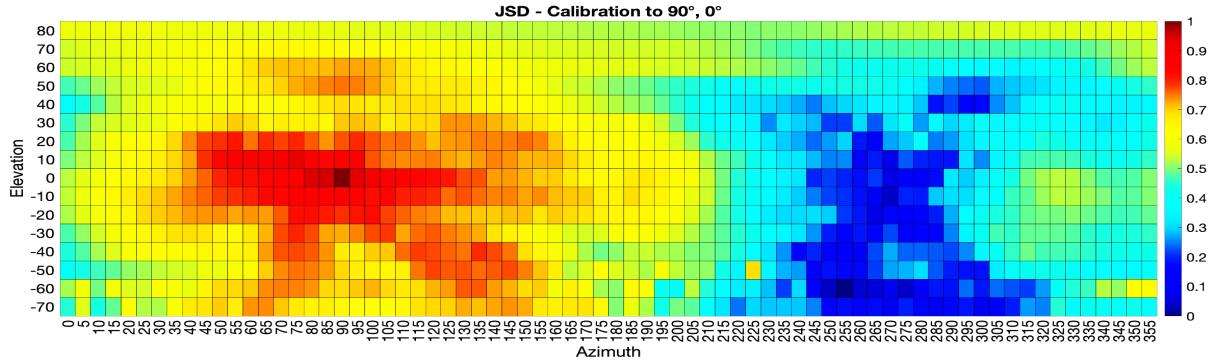

Figure 7. JSD output for initial calibration phase using test set HRIR from 90°, 0°.

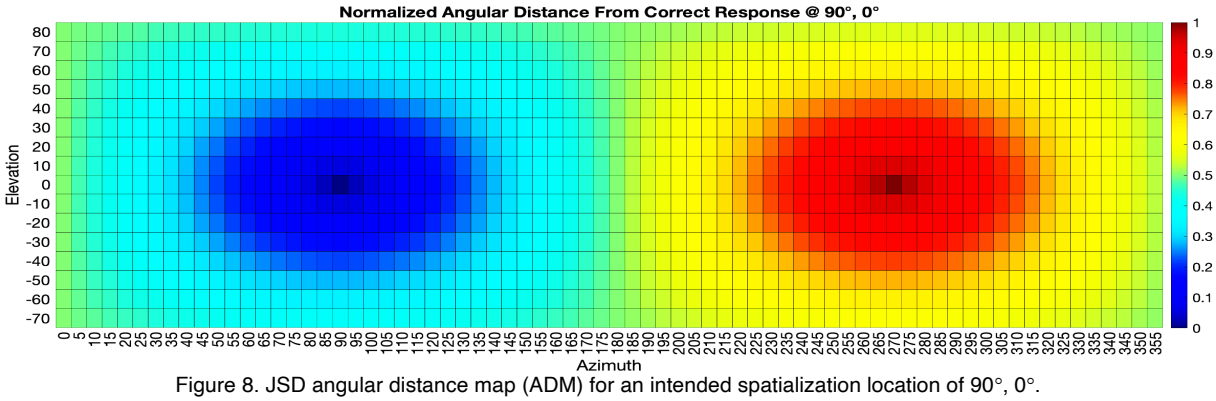

Figure 8. JSD angular distance map (ADM) for an intended spatialization location of 90°, 0°.

MOS-4 is computed as the sum over all values from the ADM and JSD product matrix. The calibration procedure presented here gives the engineer an idea for the expected lower bounds on MOS-3 and MOS-4. To compute MOS-3, first identify the index of the highest coefficient value from Fig. 7 (i.e. spatial index - 90°, 0°), then take the coefficient value in the corresponding index of the ADM shown in Fig. 8. In this instance, the value of MOS-3 is 0°. To compute MOS-4, Fig. 7 and Fig. 8 are element-wise multiplied and the resulting matrix is summed over all elements. In this instance, MOS-4 is computed as 0.2329. The expectation here is that the smaller the computed value of MOS-3, the more spatially accurate the rendering is. Similarly, the smaller the computed value of MOS-4, the more compact the virtual auditory image is.

As an additional example of the JSD processing Block, the system is run with a processed HRIR produced using VBAP on the modified cube array with intended spatialization location of (0°, 0°), and the corresponding outputs are graphically displayed in Fig. 9 and Fig. 10. The computed values for

MOS-3 and MOS-4 with a processed input of (0°, 0°) are 0° and 0.2963, respectively. This means that the intended spatialization location is equivalent to the estimated localization point, and the virtual auditory image is ~20% more diffuse than the ideal calibration reference.

4 Discussion & Conclusion

We have presented two complimentary 'Distance Computation' Blocks, one 'Perceptual Transform' Block, and demonstrated calibration and an initial application of the system. The total system outputs four metrics (MOS-1 through MOS-4) obtained from the E-N and JSD sub-components. MOS-1 and MOS-3 are intended to estimate spatialized localization precision by empirically quantifying the similarities (or distances) between the processed input response signal's spectral magnitude and the ground-truth test signal set. MOS-2 and MOS-4 are

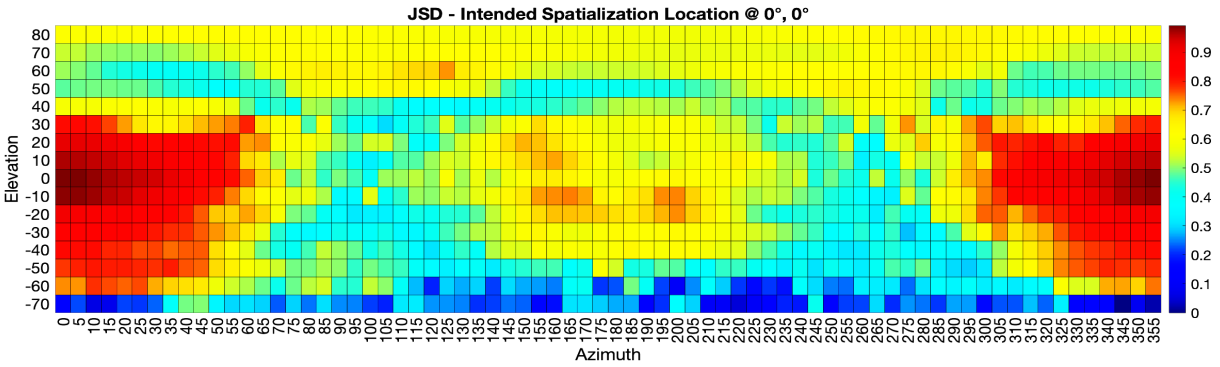

Figure 9. JSD output for processed signal input with intended spatialization location of 0°, 0°.

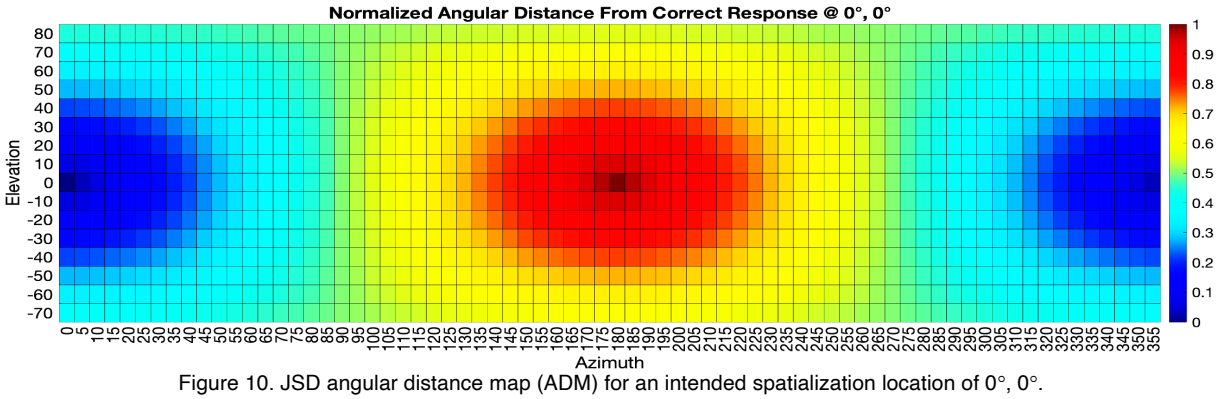

Figure 10. JSD angular distance map (ADM) for an intended spatialization location of 0°, 0°.

intended to describe the spatial variation, or "spread", of the processed response signal's spectral magnitude among the ground-truth test set. However, each produced MOS value must ultimately be verified and tuned according to professionally trained critical listeners. Additionally, in system quantification, it is often desirable to distill multiple metrics down into a single one, indicative of estimated system performance. Finding ingenuitive and psychoacoustically meaningful ways to achieve this is an active and evolving area of ongoing research.

The approach described here affords the systems engineer an ability to quantify each individual modification to the rendering pipeline and analyze the effect of said modification in a programmatic and systematic way (e.g. comparing HRTF equalization methods, Ambisonic (scene-based) vs. object-based rendering, HRTF set selection – generic vs. personal, etc.). There are also many ways to customize the proposed quantitative system in order to achieve a particular engineering requirement for any given circumstance, and a comprehensive investigation into possible sub-model component substitutions within the Perceptual Transform and Distance Computation Blocks (Fig. 2) is also an active area of ongoing research [e.g. 30, 31, 32, 33].

When HRIR individualization technology becomes ubiquitous and individual listeners entire sets are readily available, it will be quite interesting to apply them within the proposed system. Using an individualized set of HRIRs as the test signal set produces a system capable of estimating an individual listener's perceptual localization quality of experience within any given spatialization system. These systems may also be able to serve as reliable subjective listener replacements for use in wide-scale psychoacoustic investigations. However, we must reiterate the necessity of both critical listening as well as objective quantification of system performance in order to achieve a holistic approach towards spatial audio rendering evaluation.

References

[1] ITU-R Rec. BS.1387: Method for objective measurements of perceived audio quality (PEAQ), Int. Telecomm. Union, Geneva, Switzerland, 2001.

[2] ITU-T Rec. P.863: Perceptual objective listening quality assessment, Int. Telecomm. Union, Geneva, Switzerland, 2004.

[3] T. Thiede, W. Treurniet, G.A. Soulodre, Evaluation of the ITU-R objective audio quality measurement method, J. Audio Eng. Soc. 48 (3) (2000) 164-173.

[4] R. Huber, B. Kollmeier, PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception, IEEE Trans. Audio Speech Lang. Process. 14 (6) (2006) 1902-1911.

[5] ITU-T, Recommendation P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Int. Telecomm. Union, 2001.

[6] ITU-T, Rec. P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications, Int. Telecomm. Union, Geneva, Switzerland, 2004.

[7] Bocko, Mark F., and Steven Crawford. "A computational framework for objective assessment of spatial audio wavefields." *The Journal of the Acoustical Society of America* 143.3. 2018.

[8] Bocko, Mark F., Steven Crawford, and Michael Heilemann. "Prediction of Binaural Lateralization Percepts from the Coherence Properties of the Acoustic Wavefield." *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[9] Johnston, Daniel, Benjamin Tsui, and Gavin Kearney. "SALTE Pt. 1: A Virtual Reality Tool for Streamlined and Standardized Spatial Audio Listening Tests." *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[10] Sloan, Colm, et al. "Objective assessment of perceptual audio quality using ViSQOLAudio." *IEEE Transactions on Broadcasting* 63.4 (2017): 693-705.

[11] Wightman, Frederic L., and Doris J. Kistler. "The dominant role of low-frequency interaural time differences in sound localization", *The Journal of the Acoustical Society of America* 1.3, 1648-1661, 1992.

[12] Van Wanrooij, Marc M., and A. John Van Opstal. "Contribution of head shadow and pinna cues to chronic monaural sound localization." *Journal of Neuroscience* 24.17, 4163-4171. 2004

[13] Macpherson, Ewan A., and Andrew T. Sabin. "Binaural weighting of monaural spectral cues for sound localization." The Journal of the Acoustical Society of America 121.6, 3677-3688. 2007.

[14] Lopez-Poveda, Enrique A., and Ray Meddis. "A physical model of sound diffraction and reflections in the human concha." The Journal of the Acoustical Society of America100.5, 3248-3259. 1996.

[15] Zahorik, Pavel, et al. "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals." *The Journal of the Acoustical Society of America* 120.1, 343-359. 2006.

[16] Jot, Jean-Marc, Veronique Larcher, and Olivier Warusfel. "Digital signal processing issues in the context of binaural and transaural stereophony." *Audio Engineering Society Convention 98*. Audio Engineering Society, 1995.

[17] Jot, Jean-Marc, Adam Philp, and Martin Walsh. "Binaural simulation of complex acoustic scenes for interactive audio." *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[18] Jot, Jean-Marc, Véronique Larcher, and Jean-Marie Pernaux. "A comparative study of 3-D audio encoding and rendering techniques." *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[19] Jot, Jean-Marc, Scott Wardle, and Véronique Larcher. "Approaches to binaural synthesis." *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

[20] Jot, Jean-Marc, and Antoine Chaigne. "Digital delay networks for designing artificial reverberators." *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[21] Jot, Jean-Marc. "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces." *Multimedia systems* 7.1 (1999): 55-69.

[22] Audfray, Rémi, Jean-Marc Jot, and Sam Dicker. "Audio Application Programming Interface for Mixed Reality." *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[23] Lewis-Beck, Colin, and Michael Lewis-Beck. *Applied regression: An introduction*. Vol. 22. Sage publications, 2015.

[24] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the royal statistical society: series B (statistical methodology)* 67.2: 301-320. 2005.

[25] Nam, Juhan, Jonathan S. Abel, and Julius O. Smith III. "A method for estimating interaural time difference for binaural synthesis." *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[26] Nielsen, Frank. "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means." *Entropy* 21.5 (2019): 485.

[27] Armstrong, Cal, et al. "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database." *Applied Sciences* 8.11 (2018): 2029.

[28] Pulkki, Ville. "Virtual sound source positioning using vector base amplitude panning." *Journal of the audio engineering society* 45.6 (1997): 456-466.

[29] Politis, Archontis. "Microphone array processing for parametric spatial audio techniques." (2016).

[30] Trahiotis C., Bernstein L.R., Stern R.M., Buell T.N. “Interaural Correlation as the Basis of a Working Model of Binaural Processing: An Introduction”, In: Popper A.N., Fay R.R. (eds) *Sound Source Localization. Springer Handbook of Auditory Research, vol. 25*. Springer, New York, NY, 2005.

[31] Lyon, Richard F. *Human and machine hearing*. Cambridge University Press, 2017.

[32] Carney, Laurel H., and Joyce M. McDonough. "Nonlinear auditory models yield new insights into representations of vowels." *Attention, Perception, & Psychophysics* 81.4 (2019): 1034-1046.

[33] Lerud, Karl D., et al. "A canonical oscillator model of cochlear dynamics." *Hearing research* 380 (2019): 100-107.