DPLM: A DEEP PERCEPTUAL SPATIAL-AUDIO LOCALIZATION METRIC

Pranay Manocha,1⋆ Anurag Kumar,2 Buye Xu,2 Anjali Menon,2 Israel D. Gebru,2 Vamsi K. Ithapu,2 Paul Calamia2

1 Dept. of Computer Science, Princeton University, Princeton, NJ USA
2 Facebook Reality Labs Research, Redmond, WA USA

ABSTRACT

Subjective evaluations are critical for assessing the perceptual realism of sounds in audio-synthesis driven technologies like augmented and virtual reality. However, they are challenging to set up, fatiguing for users, and expensive. In this work, we tackle the problem of capturing the perceptual characteristics of localizing sounds. Specifically, we propose a framework for building a general-purpose quality metric to assess spatial localization differences between two binaural recordings. We model localization similarity by utilizing activation-level distances from deep networks trained for direction of arrival (DOA) estimation. Our proposed metric (DPLM) outperforms baseline metrics on correlation with subjective ratings on a diverse set of datasets, even without the benefit of any human-labeled training data.

Index Terms— spatial audio quality, binaural, localization, perceptual similarity, differentiable metric

1. INTRODUCTION

Perceptually realistic audio and sound-processing systems are vital for immersive multi-sensory technologies like Augmented and Virtual Reality (AR and VR) experiences. Such processing systems may include the synthesis of realistic-sounding audio, accurate spatial presentation of 3D virtual sounds, or, more broadly, high-quality rendering of virtual audio. Sound-quality evaluation tests are critical because they validate the resulting user experience and also provide necessary user feedback that drives the synthesis pipeline. Nonetheless, the inherent subjectivity in designing such tests makes it difficult to develop multi-purpose evaluation mechanisms that take various aspects of sound quality into account.

In this paper, we focus on the problem of accurate binaural presentation of sound sources in the far-field. Such presentation drives the perceptual quality of spatial audio in AR and VR. The ideal approach to characterizing binaural sound-source localization is to first synthesize the necessary sound signals and then perform a listening test via user studies. This process may be repeated hundreds of times for different combinations of source locations, which is costly and time-consuming. Further, the majority of recent audio-processing algorithms are machine learning (or deep learning) driven and rely on large labeled datasets. This makes such exhaustive listening tests impractical. The widespread use of such end-to-end systems driven by neural networks also necessitates the design of testing models that are differentiable, i.e., one can back-propagate errors from listening tests directly to the inputs. As a result, an efficient and robust objective metric that can effectively substitute for a subjective listening test is required.

Several research works have proposed objective metrics based on binaural cues like Interaural Level Differences (ILD), Time Differences (ITD) and Cross-Correlation (IACC) [1,2] to evaluate spatial audio quality. However, they suffer from various general drawbacks. First, they are sensitive to background noise - hindering their usage in diverse realistic sound synthesis scenarios. Second, they typically work well only under anechoic conditions and are not accurate in reverberant environments. Third, they don’t take into account complex scenes with multiple sources. Lastly, they assume that the two binaural signals to be compared are time-aligned and of equal length, which is not always the case. Researchers have proposed identifying the number of participating sources before using binaural cues [3], which addresses the multiple-source aspect, but the rest of the drawbacks remain.

On the other hand, one may consider adapting existing objective assessment metrics for quality of monaural signals such as PESQ [6], POLQA [7], DPAM [8] and CPDAM [9] for this task. However, since these metrics only focus on perceived quality rather than spatialization, their utility for multi-channel signals remains limited [1,10]. Some researchers have recently looked at problem-specific (e.g. audio-coding) models for objective assessment of binaural audio quality [11]. Delgado et al. [11] address the specific use-case of collapsing the stereo image to the center at low bitrates, whereas Narbutt et al. [12] compare ambisonic signals for audio codecs. These models are non-differentiable, though, and cannot be directly leveraged as a training objective for deep networks. Also, they require human-annotated datasets for training or calibration, which often are not publicly available.

We propose a framework for learning a binaural-audio similarity metric that addresses some of these issues. Specifically, we propose DPLM: a full-reference deep perceptual spatial audio localization metric that evaluates the similarity of binaural presentations in terms of localization. We begin by building binaural direction-of-arrival (DOA) deep network models that act as surrogates for localization. Given two different inputs, a simple difference of the model output layer representations between these inputs can, in principle, represent a localization metric. However, DOA estimations are typically sensitive to noise, reverberation, and sound source characteristics, which strongly affect the accuracy of localization assessment. Instead, we compute deep-feature distances [13] between the full-feature activation stacks of the DOA model to assess localization similarity between sound sources. To further improve robustness, we train DOA models with carefully designed input perturbations as data augmentations that mimic realistic environments. We show that, even in the absence of explicit perceptual training, these distances correlate well with human perceptual judgments (both via objective and subjective tests). We also show that the resulting metric generalizes even for distinct (yet related) tasks such as audio codecs, binaural reproduction from mono or multi-channel signals, etc. And since the metric is based on a deep network, it is differentiable, and can be directly leveraged as a training objective for localization and related audio and sound-source synthesis tasks.
2. THE DPLM METRIC

In this section we describe the DPLM framework (Fig 1). Given two binaural signals denoted by $x_1$ and $x_2$, our goal is to compute a distance function $D(x_1, x_2)$ that precisely characterizes the localization similarity between them. The distance function is designed to be non-negative and monotonic, thereby making it a pseudo-metric (we do not impose triangle or associative properties).

2.1. DOA Model Strategies

We begin by building a source-localization model that predicts the DOA of a given sound source. Given a binaural input, the DOA model processes both the magnitude and phase spectrograms of the two channels, and systematically outputs a framewise location estimate of the source. We evaluate two variants to ensure generalization of the proposed framework for realistic sound sources: a Static-Source model for recordings with a fixed source in the scene, and a Moving-Source model where the source can move smoothly in the scene. Note that the moving-source model would result in a finer resolution estimate of DOA.

2.2. Architecture

DPLM comprises three components (Fig 1): a feature-extraction block, a temporal aggregation block and a task-specific localization block. Both DOA variants have the same model architecture for fair comparison. The feature-extraction block maps the input features to an embedding maintaining the temporal structure of the signal. This embedding is then processed by temporal aggregation models to learn long-term dependencies. The resulting hidden representations are then fed to a task-specific localization head that outputs a location embedding per frame.

For the feature-extraction block, we evaluate a variety of different convolutional network building blocks, including a basic conv-batchnorm-maxpool block, ResNet, Squeeze-and-Excitation and Inception [14]. Inception resulted in best features, and so, we do not discuss the other structures in this paper. The 6-block Inception network consists of 64 conv filters, followed by $1 \times 1$, $3 \times 3$ and $5 \times 5$ filters, leading to $3 \times 3$ max-pooling, and a $1 \times 2$ maxpool along the frequency dimension. For the temporal aggregation block, we evaluated both Long Short-Term Memory networks (LSTMs) and Temporal Convolutional Networks (TCNs). LSTMs generalized better to unseen rooms and subjects. We used 2 bi-directional LSTM layers that output an embedding of size 64 for each time-frame.

We pose the localization task as a simple classification problem. We divide the azimuth plane ($\in (180^\circ, 180^\circ)$) into 50 equally spaced bins, and the elevation plane ($\in (-90^\circ, 90^\circ)$) into 25 equally spaced bins. However, we mainly focus on azimuthal plane localization since elevation cues are highly individualized and datasets are sparse (see also Sec 2.2). The task-specific localization head then maps the outputs from temporal aggregation to a one-hot class encoding. Note that all the layers in the blocks use BatchNorm and LeakyReLU as the activation function.

2.3. Deep-feature distance

The hidden embeddings resulting from the proposed framework are then aggregated to compute a deep-feature distance (Fig 1). Although the network is trained to predict the source location, we claim that the subspace of hidden layers contains additional useful information about estimating localization similarity. Accumulating these hidden features to compute a deep-feature distance has been shown to correlate with human perceptual judgements in some recent studies, both in machine vision [13] and audio [8,15]. They have also been shown to be effective for representation learning without the need for prior expert knowledge [13]. Given a $L$-layered network, we denote the $l^{th}$ hidden layer activations as $F_l(x) \in \mathbb{R}^{T_l \times B_l \times C_l}$, where $T_l$, $B_l$, and $C_l$ are the time resolution, frequency bands, and number of channels respectively. The distance between two audio recordings is then given by,

$$D(x_1, x_2) = \sum_l \frac{1}{T_l C_l B_l} ||F_l(x_1) - F_l(x_2)||_1.$$  (1)

2.4. Loss Functions

We train the static-source model on temporally averaged predictions, while the moving-source model uses finer, frame-level predictions. Hence its reasonable to expect that the latter model captures finer estimates of DOA. For loss function, we use the average of label-smoothed cross-entropy loss [16] and haversine distance [17].

Cross-entropy is the standard loss for classification learning with neural networks. However, label-smoothing encourages small logit gaps, which prevents model over-fitting and overconfident predictions. If $y_k$ and $p_k$ denote the target and prediction respectively ($y_k = 1$ for the correct class, 0 otherwise), and $\alpha$ denotes the smoothing parameters, the loss is given by,

$$H(y, p) = \sum_k - (y_k (1-\alpha) + \frac{\alpha}{K}) \log(p_k)$$  (2)

On the other hand, haversine distance [17] emphasizes a smooth, continuous space for DOA predictions whereas cross-entropy focuses on a discrete space. It captures the great-circle distance between any two points on a sphere and serves as a good proxy to compute distance between the predicted and ground-truth source location. Let $P = (\theta_1, \phi_1)$ and $Q = (\theta_2, \phi_2)$ be the predicted and ground truth source location from our DOA model respectively, where $\theta_1$ and $\theta_2$
We compare our approach to BAMQ, a binaural audio quality metric with annotations. We convert ambisonic formats to binaural using (BRIR) databases, listed in Table 1. The resulting pool contained The predicted source locations from the DOA model are obtained by computing a deep-feature distance (similar to Sec 2.3) between hidden layers of a pretrained deep learning model. Hence, we use two state-of-the-art binaural speech separation models - TasNet [33] and layers of a pretrained deep learning model. Hence, we use two

\[ L_{(P,Q)} = 2 \arcsin \left( \sqrt{\sin^2 \left( \frac{2\phi_1 - \phi_2}{2} \right) + \cos \left( \phi_1 \right) \cos \left( \phi_2 \right) \sin^2 \left( \frac{2\theta_1 - \theta_2}{2} \right)} \right) \]

(3)

The predicted source locations from the DOA model are obtained by computing the softmax-weighted sum of the model predictions for each frame.

3. EXPERIMENTAL SETUP

3.1. Datasets & Training

Speech recordings from the TIMIT dataset [28] are used as the source for anechoic recordings. The static-source experiments are carried out using a pool of 11 Binaural Room Impulse Response (BRIR) databases, listed in Table 1. The resulting pool contained approximately 125k BRIR pairs from 36 different rooms. For learning moving-source models, we used the Hearsay binaural audio dataset [29] which contains a total of 2 hours of paired mono and echoic binaural audio from 8 different speakers. Participants were asked to walk around a mannequin and talk (no script was used), and their position and orientation were tracked. In addition, we also used ambisonic audio data from the DCASE 2021 Challenge [30], which consists of 600 1-min long sound recordings of multiple sources with annotations. We convert ambisonic formats to binaural using the measurements from subject 2 of the ARI HRTF dataset [31].

For all cases, 3-sec audio excerpts are used for training. Phase and magnitude spectra of a 512-point DFT spectrogram are extracted from these excerpts (at 16kHz sampling rate). To ensure robustness of the metric against noise, we train DPLM with added background noise using samples from the DNS Challenge [32], spatialized using the BRIR datasets in Table 1. For learning, we use the Adam optimizer with a learning rate of $10^{-4}$ and batch size of 32. The label-smoothing parameter $\alpha$ (from Eq. 2) is 0.25. For all cases, 80% of the data is used for training and the remainder for testing.

3.2. Baselines

We compare our approach to BAMQ, a binaural audio quality metric proposed by Fleßner et al. [2]. BAMQ estimates the various binaural cues (ILD, ITD, and IACC) at frame-level and combines them using a set of learned weights to output an overall quality metric between two recordings. Further, observe that any learning model that is trained using binaural signals as inputs can, in principle, be used as a surrogate to compute a distance metric. For instance, one can compute a deep-feature distance (similar to Sec 2.3) between hidden layers of a pretrained deep learning model. Hence, we use two state-of-the-art binaural speech separation models - TasNet [33] and other alternative baselines.

### Table 1: Curated BRIR datasets for training and evaluation

| Datasets      | year | #rooms | #meas. |
|--------------|------|--------|--------|
| ADREAM [18]  | 2016 | 1      | 474    |
| AIR-1_4 [19] | 2009 | 4      | 50     |
| BRAS [20]    | 2019 | 7      | 675    |
| Huddersfield | 2019 | 1      | 1300   |
| Ilmenau [21] | 2016 | 3      | 8136   |
| IoSR [22]    | 2017 | 5      | 3641   |
| Oldenburg_IE_BTE [24] | 2009 | 5      | 296    |
| Rostock [25] | 2015 | 4      | 36288  |
| TU Berlin [26] | 2011 | 4      | 9774   |
| Salford BBC [27] | 2014 | 1      | 64800  |
| Internal Dataset | 2020 | 1    | 6      |

Figure 2: Framewise localization comparison between static and moving source DOA models. The moving trajectory in split into three intervals of constant DOA.

Figure 3: DPLM’s variation with angular distance for four fixed reference source positions. SAGRNN [34] to obtain such auxiliary localization metrics. For both these models, we compute the average of deep-feature distances across all layers except the final decoder block as alternate baselines to BAMQ.

4. RESULTS AND DISCUSSIONS

4.1. Objective evaluations

We first evaluate the static-source and moving-source models for localization errors on a held-out set of sound sources from TIMIT, spatialized using a held-out set of BRIRs. The best performing static-source model produced a root mean square error (RMSE) of 13.2° in azimuth front-back folded, i.e., reflected about the coronal plane to discount front-back confusions. The moving-source model produced an RMSE of 8.4° confirming that it leads to more accurate DOA estimates.

Fig 2 shows an example of a framewise comparison between static-source and moving-source models. We observe that the framewise predictions from the moving-source model (blue curve) closely follow the actual trajectory of the source (black curve). As expected, the static-source model (orange dotted curve) is not accurate at the frame level for tracking moving objects. On the other hand, the prediction improves (red curve) when localization is computed independently for each interval, after splitting the moving trajectory into various intervals of constant DOA (shown by the three intervals in Fig2). All these observations are expected, and overall,
the results show that the continuous temporal tracking information available to the moving-source model helps improve the frame-level predictions, leading to fewer localization errors in general.

To verify DPLM’s sensitivity to increasing angular distance, Fig. 3 shows the metric’s distance between a fixed reference, and a moving test source for four different source positions. We see that the absolute distance values generally increase with increasing angular distance across all four source positions, indicating that DPLM obeys the general trend quite well. To quantify this trend, Table 3 shows the Spearman’s rank order correlation (SC) between the output of DPLM and angular distance between two sources across subjects and (anechoic/echoic) listening conditions. We see that both our models (static-source and moving-source) outperform all baselines. Surprisingly, even the pre-trained models have a non-trivial correlation with angular distance, suggesting that deep-feature distances across these models serve as a good proxy for assessing localization differences between recordings.

Results for the correlations with subjective ratings are in Table 2. Overall, our proposed metric achieves the best performance across all datasets. Firstly, DPLM’s correlation is stable with changes in stimuli (shown by P1, P1’ and P4). This shows the generalization power of deep-feature distance metrics, and their ability to capture attributes across speech, music, noise etc. Furthermore, the two deep network baselines (TasNet and SAGRNN) trained on an unrelated task (binaural source separation) outperform BAMQ on most datasets. This suggests that deep-feature distances transfer well even across unrelated tasks, and are able to model low-level perceptual similarity well. However, absolute correlation values are lower for P3 showing that the metric is not robust enough to capture subtle differences driven by headphone equalization (some of which are very close to JNDS). Secondly, the moving-source model performs better than the static-source model on most tasks, following a similar trend as shown in Table 3 earlier. Hence, framework optimization for localization also seems to improve subjective ratings. Third, the trends with P2, P3 and P4 suggest that DPLM (and the two deep network pre-trained baselines) are robust to non-time-aligned data. Lastly, since DPLM is trained and tested under realistic, echoic, conditions, we can clearly see its better generalization power compared to BAMQ (which is learned under anechoic environments).

Recall that one can characterize azimuth localization by utilizing binaural cues such as ITD and ILDs. However, elevation localization is quite challenging because of the subject-specific influence of monaural spectral cues. Further, lack of a wide range of elevation angles in publicly available BRIR datasets also limits building and evaluating robust models. We also observed similar trends in our analysis (not shown), with high error for elevation localization. The proposed metric performed almost the same as a simple spectral subtraction, suggesting that it does not capture elevation cues well.

5. CONCLUSIONS AND FUTURE WORK

We present DPLM, a full-reference, general purpose, differentiable perceptual objective metric to assess spatial localization differences between two binaural recordings. We show that deep-feature distances obtained from DOA models correlate well with human ratings of localization similarity across a variety of datasets. This is achieved without any perceptual training or calibration. In the future, we would like to extend this metric to improve elevation localization, as well as improve performance for recordings that have subtle differences. One can also explore the utility of these differentiable metrics in deep learning based binaural speech enhancement and synthesis methods.
6. REFERENCES

[1] S. Kampf, J. Liebetrau, et al., “Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio,” Journal of the AES, october 2010.

[2] J.-H. Fleßner, R. Huber, et al., “Assessment and prediction of binaural aspects of audio quality,” Journal of the AES, vol. 65, no. 11, 2017.

[3] J.-H. Seo, S. B. Chon, et al., “Perceptual objective quality evaluation method for high quality multichannel audio codecs,” Journal of the AES, vol. 61, no. 7/8, 2013.

[4] M. Takanen and G. Lorho, “A binaural auditory model for the evaluation of reproduced stereophonic sound,” in AES: Applications of Time-Frequency Processing in Audio, 2012.

[5] M. Schäfer, M. Bahram, et al., “An extension of the PEAQ measure by a binaural hearing model,” in ICASSP, 2013.

[6] A. W. Rix, J. G. Beerends, et al., “Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment,” in ICASSP, 2001.

[7] J. G. Beerends, C. Schmidmer, et al., “Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality,” Journal of the AES, vol. 61, no. 6, 2013.

[8] P. Manocha, A. Finkelstein, et al., “A differentiable perceptual audio metric learned from just noticeable differences,” Interspeech, 2020.

[9] P. Manocha, Z. Jin, et al., “CDPAM: Contrastive learning for perceptual audio similarity,” ICASSP, 2021.

[10] R. Conetta, T. Brookes, et al., “Spatial audio quality perception (part 1): Impact of commonly encountered processes,” Journal of the AES, vol. 62, no. 12, 2015.

[11] P. M. Delgado and J. Herre, “Objective assessment of spatial audio quality using directional loudness maps,” in ICASSP, 2019.

[12] M. Narbutt, J. Skoglund, et al., “Ambiual: Towards a quality metric for headphone rendered compressed ambisonic spatial audio,” Applied Sciences, vol. 10, no. 9, 2020.

[13] R. Zhang, P. Isola, et al., “The unreasonable effectiveness of deep features as a perceptual metric,” in CVPR, 2018.

[14] C. Szegedy, W. Liu, et al., “Going deeper with convolutions,” in CVPR, 2015.

[15] F. G. Germain, Q. Chen, et al., “Speech denoising with deep feature losses,” Interspeech, 2019.

[16] C. Szegedy, V. Vanhoucke, et al., “Rethinking the inception architecture for computer vision,” in CVPR, 2016.

[17] Z. Tang, J. D. Kanu, et al., “Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks,” Interspeech, 2019.

[18] F. Winter, H. Wierstorf, et al., “Database of binaural room impulse responses of an apartment-like environment,” in AES Convention 140, 2016.

[19] M. Jeub, M. Schafer, et al., “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in ICDSP, 2009.

[20] L. Aspöck, F. Brinkmann, et al., “BRAS-Benchmark for room acoustical simulation,” 2020.

[21] B. I. Bacila and H. Lee, “360° binaural room impulse response (BRIR) database for 6df6 spatial perception research,” in AES Convention 146, 2019.

[22] C. Mittag, M. Böhme, et al., “Dataset of KEMAR-BRIRs measured at several positions and head orientations in a real room,” Dec. 2016.

[23] J. Francombe, “IoSR listening room multichannel BRIR dataset.”

[24] H. Kayser, S. D. Ewert, et al., “Database of multichannel in-ear and behind-the-head related and binaural room impulse responses,” EURASIP Journal on ASP, 2009.

[25] V. Erbes, M. Geier, et al., “Database of single-channel and binaural room impulse responses of a 64-channel loudspeaker array,” in AES Convention 138, 2015.

[26] H. Wierstorf, M. Geier, et al., “A free database of head related impulse response measurements in the horizontal plane with multiple distances,” in AES Convention 130, 2011.

[27] D. Satongar, Y. W. Lam, et al., “Measurement and analysis of a spatially sampled binaural room impulse response dataset,” in 21st International Congress on Sound and Vibration, 2014.

[28] J. S. Garofolo, L. F. Lamel, et al., “DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1,” NASA Technical Report, vol. 93, p. 27403, 1993.

[29] A. Richard, D. Markovic, et al., “Neural synthesis of binaural speech from mono audio,” in ICLR, 2021.

[30] A. Politis, S. Adavanne, et al., “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” in DCASE2020.

[31] P. Majdak, “ARI HRTF database,” Dec. 2017. [Online]. Available: https://www.oew.ac.at/isl/das-institut/software/hrtf-database

[32] C. K. Reddy, E. Beyrami, et al., “The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework,” Interspeech2020.

[33] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” ACM TASLP, 2019.

[34] K. Tan, B. Xu, et al., “SAGRNN: Self-attentive gated rnn for binaural speaker separation with interaural cue preservation,” IEEE SPL, 2020.

[35] Z. Ben-Hur, D. L. Alon, et al., “Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs,” ACM TASLP, 2021.

[36] T. Lühbeck, H. Helmholz, et al., “Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data,” Journal of the AES, vol. 68, no. 6, 2020.

[37] I. Engel, D. L. Alon, et al., “The effect of generic headphone compensation on binaural renderings," in AES Conference on Immersive and Interactive Audio, 2019.

[38] T. Rudzki, I. Gomez-Lanzaco, et al., “Perceptual evaluation of bitrate compressed ambisonic scenes in loudspeaker based reproduction," in AES Conference on Immersive and Interactive Audio, 2019.