2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 978-1-6654-4870-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/WASPAA52581.2021.9632781

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 Calamia*[2]

[1] Dept. of Computer Science, Princeton University, Princeton, NJ USA
[2] Facebook Reality Labs Research, Redmond, WA USA
pmanocha@cs.princeton.edu, {anuragkr90, xub, aimenon, idgebru, ithapu, pcalamia}@fb.com

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) [1–4] 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 [3,5], 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 CDPAM [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 [3–5, 11, 12]. 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.

⋆ This work was performed during an internship at Facebook Reality Labs Research

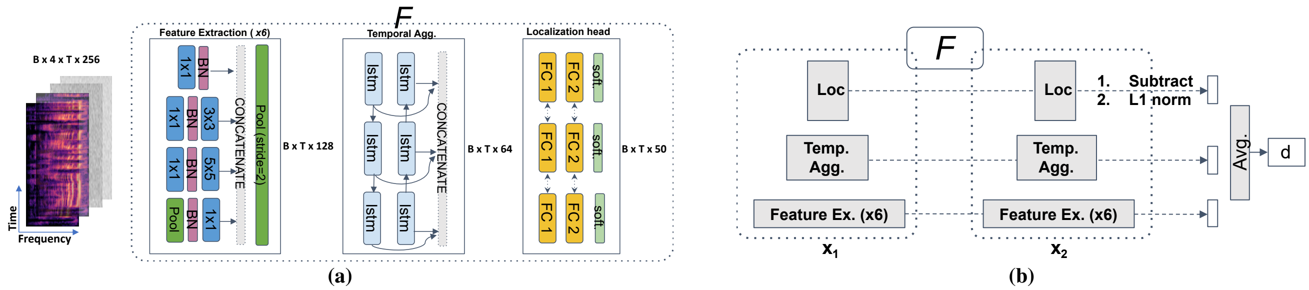

Figure 1: **DPLM architecture**: (a) We first train a source localization model F to estimate framewise *DOA*, then (b) use the extracted *deep-features* to compute a distance $D(x_1, x_2) = d$ between the two binaural signals x_1 and x_2

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 1a): a *feature-extraction* block, a *temporal aggregation* block and a task-specific *localization block*. Both DOA variants have the same architecture for fair comparison. The feature-extraction block maps the input to an embedding maintaining the temporal structure of the signal. This embedding is processed by temporal aggregation models to learn long-term dependencies. The resulting representations are fed to task-specific localization heads that output a location per frame.

For the feature-extraction block, we evaluated 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 1x1, 3x3 and 5x5 filters, leading to 3x3 max-pooling, and a 1x2 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 focus on azimuthal plane localization since elevation cues are highly individualized and datasets are sparse (see also Sec 4.2). 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 1b). 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,

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

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 the 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 α denotes the smoothing parameters, the loss is given by,

$$H(\mathbf{y}, \mathbf{p}) = \sum_{k=1}^{K} -(y_k(1-\alpha) + \frac{\alpha}{K}) \log(p_k) \quad (2)$$

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 θ_1 and θ_2 are azimuth angles, and ϕ_1 and ϕ_2 are the elevation angles (all in radians), then the distance between P and Q is:

$$L_{(P,Q)} = 2 \arcsin\left[\sqrt{\sin^2(\tfrac{\phi_1 - \phi_2}{2}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\tfrac{\theta_1 - \theta_2}{2})}\right] \quad (3)$$

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

Datasets	year	#rooms	#meas.
ADREAM [18]	2016	1	474
AIR_1_4 [19]	2009	4	50
BRAS [20]	2019	7	675
Huddersfield [21]	2019	1	1300
Ilmenau [22]	2016	3	8136
IoSR [23]	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

Table 1: Curated BRIR datasets for training and evaluation

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 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 α (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 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.

Note that we do not include any specific baselines using conventional DOA models. This is because the main focus of our work is to design an audio quality metric, which is why BAMQ is an appropriate baseline. However, note that BAMQ is trained and tested under anechoic conditions, unlike DPLM which is trained with echoic BRIRs.

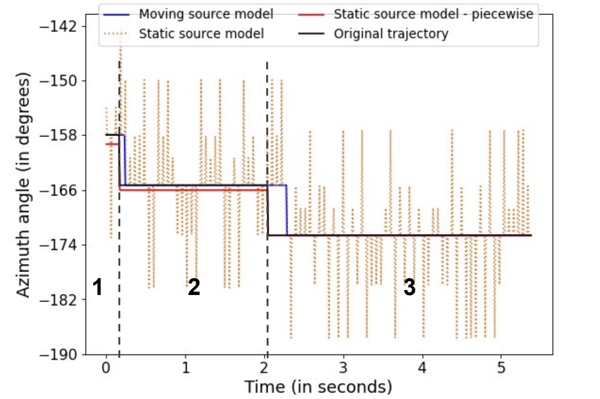

Figure 2: ***Framewise* localization** comparison between *static* and *moving* source DOA models. The *moving* trajectory is split into three intervals of *constant* DOA.

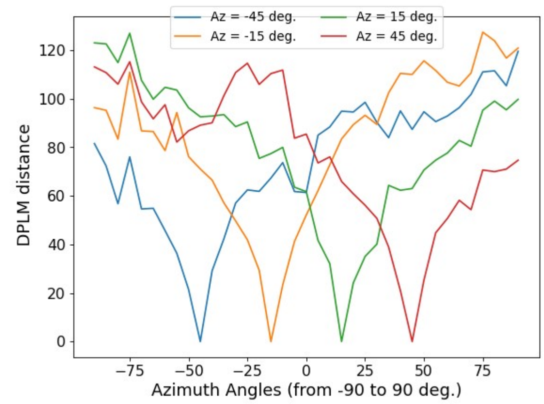

Figure 3: DPLM's variation with angular distance for four *fixed* reference source positions.

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**$^\circ$ 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**$^\circ$ 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 Fig 2). 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

Type	Name	P1			P1’		P2	P3			P4					
		Speech	**Castanets**	**Guitar**	**Speech**	**Castanets**	**Music**	**Speech**	**Pink Noise**	**Guitar**	**Pink Noise**	**Vocals**	**Castanets**	**Glocken**	**EM**	**AM**
Pre-trained	**TASNET**	0.65	0.48	0.20	0.65	0.35	0.29	0.19	0.20	**0.10**	0.45	0.01	0.20	0.12	0.61	0.69
	SAGRNN	0.72	0.61	0.21	0.65	0.40	0.37	0.20	0.24	0.07	0.45	0.14	0.36	0.19	0.61	0.72
DOA Models	**static-source**	0.89	0.91	0.85	0.82	**0.94**	**0.45**	0.59	**0.33**	0.07	**0.53**	**0.62**	0.36	0.45	0.61	0.79
	moving-source	**0.94**	**0.94**	**0.94**	**0.83**	**0.94**	**0.45**	**0.69**	0.22	0.06	**0.53**	0.61	**0.42**	**0.47**	**0.67**	**0.83**
Baseline	**BAMQ**	0.03	0.83	0.09	0.52	0.77	-0.17	0.42	0.21	0.08	-0.02	0.36	0.11	-0.05	0.23	0.18

Table 2: **Subjective evaluation**: Models include: Pre-trained models, our DOA models (including *static-source* and *moving-source* models), and BAMQ, as baseline. Spearman Correlation (SC). ↑ is better.

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. Note here that we do not convert BAMQ into a DOA prediction model. We merely correlate the distance from BAMQ with angular distance between the two sources. Ideally, a larger angular distance should have a higher distance from BAMQ .

	BAMQ	TASNET	SAGRNN	*static-source*	*moving-source*
Localization	0.16	0.24	0.67	0.82	**0.86**

Table 3: **Objective evaluation**: Correlation with *angular distance*. Models include: Pre-trained models, our DOA models (including *static-source* and *moving-source* models), and BAMQ as baseline. Spearman Correlation (SC). ↑ is better.

4.2. Subjective evaluations

We now use previously published diverse third-party studies to verify that our trained metric correlates well with subjective ratings of their tasks. We compute the correlation between the proposed model's predicted distance with the publicly available subjective ratings. These correlation scores are evaluated per condition (averaging samples per condition). The datasets used are:

1. **Bilateral Ambisonics [35] (P1 and P1')**: This compares the standard and bilateral spatial audio reproduction methods across various spherical harmonic orders to assess overall quality. It uses various stimuli including speech, castanets and guitar. There are two versions (denoted by P1 and P1'), each with different subjects, training sessions and different spherical-harmonic orders (distortions are mainly spatial).
2. **Spherical Microphone Array [36] (P2)**: This is designed to compare audio quality improvements across algorithms for binaural rendering of spherical microphone array signals using music as stimuli. It provides an overall quality rating, with 96 variations of test signals per subject. The pairs of recordings to be compared are not time-aligned and can be of different lengths (distortions are mainly spatial).
3. **HpEQ [37] (P3)**: This data is from headphone equalization (HpEQ) study across generic and individualized BRIRs, with individualized, generic or no headphone equalization using speech, pink noise and guitar sounds for stimuli. We have an overall quality rating, and pairs of recordings may contain very subtle differences (recordings are also not time-aligned, and distortions are mainly spatial).
4. **Bitrate Compr. Ambis. [38] (P4)**: This comes from assessment of the degree of timbral distortions introduced by compression at different Ambisonic orders (1st, 3rd and 5th) across various modalities including simple scenes (vocals, castanets, glockenspiel, pink noise) and complex scenes (EM: electronic music and AM: acoustic music). Similar to P3, we have overall quality ratings, and recordings are not time-aligned (distortions consist of spatial and monaural distortions).

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, frame-wise 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. Finally, note that we compare with overall audio quality across all 4 datasets, which might suggest why BAMQ performs worse, since it was designed to quantify spatial distortions. However, BAMQ performs worse across datasets that mainly consist of spatial distortions (*P1*, *P1'*, *P2* and *P3*). This suggests the better generalization power of DPLM compared to BAMQ, given that our model is trained under realistic, echoic conditions.

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. We would also like to focus on more complex, realistic audio scenes consisting of more than a single source. 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.*, "Ambiqual: 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 6dof 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," 2016 - https://doi.org/10.15126/surreydata.00813511.

[24] H. Kayser, S. D. Ewert, *et al.*, "Database of multichannel in-ear and behind-the-ear 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.oeaw.ac.at/isf/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. Lübeck, 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.