INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Spatialization Quality Metric for Binaural Speech

*Pranay Manocha*[1*], *Israel D. Gebru,*[2] *Anurag Kumar,*[2] *Dejan Markovic,*[2] *Alexander Richard*[2]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]Meta Reality Labs Research, USA

pmanocha@princeton.edu, {idgebru,anuragkr90,dejanmarkovic,richardalex}@meta.com

## Abstract

In spatial-audio enabled systems, evaluating the quality of spatialization is an essential process. This paper proposes a new objective metric to measure the spatialization quality (SQ) between any pair of binaural signals while being agnostic to speech content and signal duration. We formulate SQ as a metric learning problem and compute deep-feature distance on embeddings learned using triplet loss and multi-task learning with direction-of-arrival and binaural speech synthesis as auxiliary tasks. We show the robustness of our model on localization in (un)seen contexts, monotonicity with increasing angular distance, content in-variance and retrieval performance. Experiments show that our metric correlates well with publicly available subjective ratings, and it yields improvements when used as a differentiable loss in a binaural speech enhancement system.

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

## 1. Introduction

Many applications that (re)produce audio signals require evaluating audio quality. To date, human judgement via listening tests conducted in a controlled environment, also known as subjective evaluations, is the most reliable method for assessing quality, e.g., MUSHRA [1]. Nonetheless, these tests are time-consuming and expensive to conduct. In general, they cannot be easily carried out at system development stages, therefore preventing us from evaluating novel research ideas and frameworks quickly. To address this challenge, numerous computational tools that are more accessible and practical have been developed in the last few decades. These methods, also known as objective metrics, are typically developed for assessing a specific audio attribute or application domain in mind [2, 3].

In the domain of spatial audio processing, evaluating audio quality remains an open and challenging problem because one must assess quality across multiple attributes [4], including *perceptual audio quality* (PAQ) and *spatialization quality* (SQ). PAQ refers to measuring audio material distortion and artifacts affecting a specific signal, e.g., speech quality, intelligibility, etc. On the other hand, SQ measures how accurately individual sound sources are positioned in a virtual space to be accurately localized in the rendered audio, e.g., when listening over headphones [5], over loudspeaker systems [6]. In this paper, however, we focus on measuring the SQ of binaural signals.

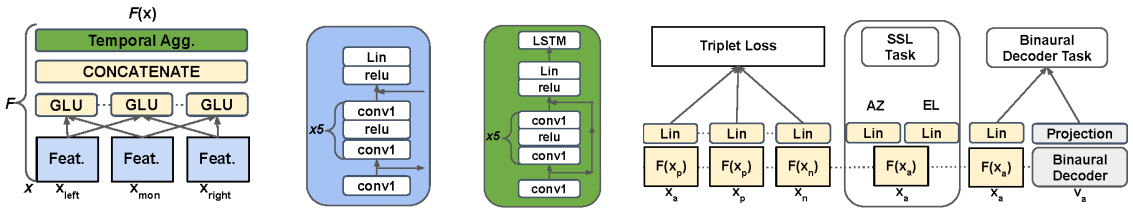
The earliest research on SQ employs human binaural hearing-inspired sound source localization models as metrics [4, 7]. These models rely heavily on signal processing pipelines and carefully engineered design choices exploiting in-domain knowledge of binaural hearing and psycho-acoustics [8, 9]. Several other methods proposed extending the standard monaural audio perceptual quality metrics such as PEAQ [10] and POLQA [11] as SQ metrics [12–15]. The work in [14, 16] proposed to extend PEAQ to multiple channels by adding binaural hearing models. Similarly, the work in [13] proposes to incorporate known binaural auditory cues [17] such as Interaural Level and Time Differences (ILD,ITD) and Interaural Cross-Correlation (IACC), which can be extracted from the left and right channel of binaural signals. Binaural cues preserve some information about SQ; however, in most of these methods, different auditory models were used as monaural or binaural cues, and thus their interdependence and interaction is not well modeled. Additionally, the majority of these methods are not versatile [13, 18, 19]. They require either clean reference signals (*i.e. full-reference*) or signals with identical content (*i.e. matched-reference*) to compare with, and thus cannot be used when paired clean reference is unavailable. Furthermore, they presume that test signals are created in relation to the reference, i.e. the two signals are time-aligned and of equal length, which may not always be the case during testing.

More recently, machine learning has provided robust, rapidly re-trainable, and easily expandable solutions that can be used in many other spatial audio-related problems. The work in DPLM [18] proposed a full-reference spatialization metric that evaluates the similarity of binaural signals based on source localization. This work uses a pre-trained Direction-of-arrival (DoA) model and computes deep-feature distances between the outputs of intermediate layers. Similarly, SAQAM proposed in [20] jointly assesses PAQ and SQ for non-matching reference signals; it combines these two tasks using a multi-task learning framework [21] to leverage the useful information contained in related tasks and improve generalization performance. Note that both DPLM and SAQAM use pre-trained DoA models with coarsely discretized azimuth and elevation directions (e.g., $10^\circ$) as classification/regression targets. It is unclear whether the DoA models provide reliable localization estimates. Even though the metrics are differentiable and do not require a clean reference or time-aligned signals, they only work on signals with a sampling rate of 16kHz; limiting their utility as loss functions in high-fidelity audio tasks that use high sampling rates, e.g., 48 kHz.

To tackle some of the aforementioned challenges and drawbacks, we introduce a novel metric to evaluate SQ of binaural speech signals and call it **SQM-BS** (***S**patialization **Q**uality **M**etric for **B**inaural **S**peech*). Inspired from work in [18, 20], we formulate the quality metric as a deep metric learning (DML) problem and compute similarity on embedding vectors extracted from 2D direction-of-arrival (DoA) estimation

*Work done during internship at Meta Reality Labs Research.


10.21437/Interspeech.2023-832




(a) (b) (c) (d)

Figure 1: *Details of the **SQM-BS**. (a) Network architecture, (b) The feature extraction block, (c) The temporal aggregation blocks, (d) Training framework with triplet loss and two additional tasks including Sound source localization (SSL) and Binaural Speech Synthesis embeddings. Note that blocks connected through dotted lines share weights.*

model. We followed a triplet loss architecture and trained the DoA together with a recently proposed binaural speech synthesis model [22] in a multi-task learning setting. This allows us to learn not only robust embeddings but also content-invariant representations that make our model more versatile. Consequently, we neither require full-reference nor matched-reference signals. The use of DML gives us the flexibility to map audio to an embedding vector which can be computed once for a static sound source or on a frame-by-frame basis for a moving source. As a result, our model does not require time-aligned signals. Moreover, in order to make our model robust to echoic conditions, we augmented our speech-based training data with realistic room reverb and noise data. We evaluate our formulation using a number of multi-task learning strategies and compare different models in terms of localization errors in (un)seen contexts, monotonicity with increasing angular distance, content in-variance, retrieval performance, correlation to subjective ratings across four tasks, as well as on a newly collected 2AFC (2-alternative forced choice) dataset. Lastly, since our model is differentiable, it can be used as a loss function to train other tasks. We showed that training an existing binaural speech enhancement system with our metric yields significant improvements.

## 2. The SQM-BS Framework

The SQM-BS network architecture and training framework are shown in Figure 1. The network takes a binaural signal (left and right channels) and its corresponding mono recording as inputs. First, the binaural-mono pair from a test recording and its corresponding reference recording are separately fed into the network. The output of the last layer is calculated and referred to as *feature map*. Following that, spatialization quality metrics are computed based on similarities between the feature maps of the test recording and that of the reference recording.

Let $\boldsymbol{x}_A$, $\boldsymbol{x}_B$ and $\boldsymbol{x}_c$ denotes test recordings; each containing binaural-mono pair. Let $\mathbf{d}(\cdot,\cdot)$ denote the similarity distance function calculated by SQM-BS. It has the following properties: (1) Non-Negative: $\mathbf{d}(\boldsymbol{x}_A, \boldsymbol{x}_B) \geq 0$; (2) Monotonic: if $\mathbf{d}(\boldsymbol{x}_A, \boldsymbol{x}_B) \geq \mathbf{d}(\boldsymbol{x}_C, \boldsymbol{x}_B)$, then $|\mathcal{L}(\boldsymbol{x}_C) - \mathcal{L}(\boldsymbol{x}_B)| \leq |\mathcal{L}(\boldsymbol{x}_A) - \mathcal{L}(\boldsymbol{x}_B)|$, where $\mathcal{L}$ denotes the sound source location; (3) Indiscernibility of Identicals: $\mathbf{d}(\boldsymbol{x}_A, \boldsymbol{x}_A)$ has very small (close to zero) scores; and (4) In-variant to content: $\mathbf{d}(\boldsymbol{x}_A, \boldsymbol{x}_B)$ does not depend on speech content in $\boldsymbol{x}_A$ or $\boldsymbol{x}_B$. The first two properties could be trivially achieved by design. To effectively model the last two, we devise triplet learning and multi-task learning approach. As an outcome, we created an embedding space in which feature maps are content in-variate representations of the spatialization property of binaural signals. Details are sequentially presented in the following sections.

### 2.1. Triplet Learning

Recently, triplet learning has attracted increasing attention as a popular deep metric learning (DML) method and has shown considerable potential in constructing task-specific distance metrics from (weakly-) supervised data, as it can help models learn a measure of distance by the notion of similarity and dissimilarity. We define triplets as a set $\mathcal{T} = \{t_i\}_{i=1}^{N}$ where a triplet $t_i = \{\boldsymbol{x}_a^i, \boldsymbol{x}_p^i, \boldsymbol{x}_n^i\}$ contains $\boldsymbol{x}_a$ as the anchor sample, $\boldsymbol{x}_p$ as the positive sample, and $\boldsymbol{x}_n$ as the negative sample. We define triplet loss [23] as:

$$L(t) = max\{0, \mathbf{d}(\mathbf{f}(\boldsymbol{x}_a), \mathbf{f}(\boldsymbol{x}_p)) - \mathbf{d}(\mathbf{f}(\boldsymbol{x}_a), \mathbf{f}(\boldsymbol{x}_n)) + \delta\}, \quad (1)$$

where $\mathbf{f}$ is the function approximated by the model weights, $\delta$ is a margin value to prevent trivial solutions, and $\mathbf{d}(\mathbf{f}(\boldsymbol{x}_1), \mathbf{f}(\boldsymbol{x}_2))$ is the cosine similarity between the embeddings of the two input recordings ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$). Note that, in our case, we design triplets such that the anchor and positive samples have closer localization than the anchor and negative.

### 2.2. Multi-task learning (MTL)

MTL has been shown to be useful in learning robust and generalizable representations in a variety of speech applications [24]. It leverages useful information contained in multiple related tasks to help improve the generalization performance of each task. With this in mind, we introduce two new tasks to enforce the proposed model to learn content-invariant but sound source location-dependent feature maps.

**SSL:** We trained a sound source-localization (SSL) model that predicts the direction of arrival (DOA) of a given binaural audio recording. We used the SSL model as used in [18, 20]. This task takes the feature maps from a pre-trained SQM-BS model and learns a projectible mapping to a discrete space of sound source locations.

**Binaural-Decoder:** This task makes use of pre-trained embeddings from a pre-existing binaural speech synthesis [22]. We learn a projection from our model feature maps to the binaural synthesis module's adapter block outputs. The adapter module take view vectors (aka 6 DoF coordinates) as input. By optimizing the feature maps to match the outputs of the adaptor, we ensure that our model is independent of speech content but dependent on view (position) information.

**Combined:** This model is our proposed full **SQM-BS** metric model. It combines the previous two tasks together with the triplet learning framework.

### 2.3. Architecture

The network architecture of our model is shown in Figure 1. It consists of three components: feature-extraction block, temporal aggregation block and task-specific heads. The feature-extraction block (Figure 1(b)) is a 1D ResNet-styled architecture without stride to preserve information and maintain temporal consistency. The temporal aggregation block (Figure 1(c)), contains 2 bi-directional LSTM layers that output a feature map of size 64 for each time frame. The architectures for task-specific heads are task-dependent. For the task of triplet learning, the temporally aggregated feature maps are transformed through a linear layer to output an embedding vector of dimension 16; the triplet loss is then applied to this embedding.

### 2.4. Loss functions

For the triplet learning formulation, we use cosine distance as a distance function with a margin $\delta$ (see Equation 1). For the SSL subtask, we use the label-smoothed Cross-Entropy (CE) loss as used in [18, 20]. For the binaural synthesis subtask, we used the cosine distance between the output of our model and the target binaural synthesis embedding as a loss function.

### 2.5. SQM-BS as a metric loss

We use deep feature distances as a proxy similarity metric. These have been found to be robust to imperceptible differences, and correlate well with human perception of similarity judgments [18, 20, 25]. Given a L-layered network, the output of $l^{th}$ hidden layer is $\mathbf{f}_l(x) \in \mathbb{R}^{\mathrm{T}_l \times \mathrm{C}_l}$, where $\mathrm{T}_l$ and $\mathrm{C}_l$ are the time resolution and number of channels respectively. The distance between two audio recordings is then given by:

$$\mathbf{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_l \frac{1}{\mathrm{T}_l \mathrm{C}_l} ||\mathbf{f}_l(\boldsymbol{x}_1) - \mathbf{f}_l(\boldsymbol{x}_2)||_1. \quad (2)$$

## 3. Experimental Setup

### 3.1. Datasets and Training

Since there is no publicly available binaural audio dataset with paired mono and source location data required to train and validate our proposed framework, we collected a novel binaural audio dataset. We re-recorded 42 hours of speech from the VCTK corpus [26] using 108 3Dio binaural microphones placed in a non-anechoic room. 96 microphones were placed at various height levels around a circular recording area, and the remaining 12 microphones were placed at the center. We played speech signals from VCTK corpus over a small hand-held loudspeaker which was carried by a person walking around the room. The 3D position and orientation of the loudspeaker as well as the binaural microphones were tracked using Optitrack system. With this setup, we recorded 42 hours of binaural audio data, covering a distance of 4.6m horizontally and 2.4m vertically. The signal sent to the loudspeaker is considered as the mono source. In total, we have collected $42 \times 108 = 4526$ hours of binaural audio data. Let $\mathcal{D}_1$ denotes this dataset.

Our model takes the raw binaural signals and their mono counterpart as inputs. We found that adding the mono signal helps the model to perform better (see Section 4.3 for comparisons with binaural-only input). For the triplet learning, the inputs to be model $(\boldsymbol{x}_a, \boldsymbol{x}_p, \boldsymbol{x}_n)$ are created by sampling *different* recordings from $\mathcal{D}_1$. For training, we use the Adam optimizer with a learning rate of $10^{-4}$ and batch size of 32. The label-smoothing parameter $\alpha$ is 0.25. The margin $\delta$ is set to 1.

### 3.2. Evaluation and Baselines

We compared our model with four existing methods. BAMQ [13] is a binaural audio quality metric that estimates quality from binaural cues such as ILD, ITD, and IACC. GPSM+BMFD [19] combines monaural and binaural psychoacoustic models in a multi-stage processing step to estimate quality. DPLM [18] and SAQAM [20] are deep learning-based quality metrics.

To benchmark our model on unseen data, we augment $\mathcal{D}_1$ dataset using a pool of 11 publicly available Binaural Room Impulse Response (BRIR) databases including Huddersfield [27], Ilmenau [28], and IoSR [29]. We used speech recordings from the TIMIT [30] dataset as the source for anechoic recordings, and pyroomacoustics toolkit [31] to synthesize the binaural recordings. This dataset is denoted as $\mathcal{D}_2$.

Additionally, we benchmark the performance of the SSL model used in our framework by comparing it with that of DPLM [18] and SAQAM [20]. We compared the generalization capabilities of two variants across unseen datasets. The first approach is a *classification* model in which the output/target is discretized into equally spaced azimuth and elevation directions. The second is a *regression* model where the model's output is the raw azimuth and elevation values. We trained all methods and their variants on recordings with a sampling rate of 48kHz.

## 4. Results

### 4.1. Objective Evaluations

**Localization**: We evaluate the models for localization errors (both az and el) on a held-out set of recordings from $\mathcal{D}_1$ and $\mathcal{D}_2$. We report the root-mean-square errors (RMSE) between the predicted localization and the ground truth estimates (in degrees). The results are shown in Table 1. The RMSE scores of our models are low compared to other baseline methods.

**Monotonicity**: We create a test dataset having non-matched audio recordings at increasing angular distances, and compute Spearman rank order correlation (SC) between the distance from the metric and angular distance. The results are reported in Table 1. Our combined model has the highest SC score suggesting a strong correlation between increasing angular distance and model distance - indicative of monotonicity.

**Content in-variance**: To evaluate robustness to content variations, we created two groups of test datasets: one consists of pairs of recordings with the same localization but different speech content; the other contains pairs of recordings with different localizations and speech. We calculate the common area (CA) between normal distributions computed across the models' output from the two groups. The smaller the common area, the greater the "difference" between the two groups. As shown in Table 1, our combined framework has the lowest CA score, suggesting in-variance to content.

**Retrieval**: We evaluate retrieval performance to measure the quality of the top-K items in the ranked list. We divide the entire localization space into 20 groups, each group consisting of 100 recordings with the same source location, but different speech content. We took randomly selected queries and calculate the number of correct class instances in the top-K retrievals, and report mean precision@K ($\mathrm{MP}_k$) over test queries. Most models have high precision suggesting that embeddings capture SQ attributes (see Table 1).

### 4.2. Subjective Evaluations

We validate the correlation of our trained metric with subjective ratings using a dataset of publicly available subjective ratings.

| Name | Models | RMSE on $\mathcal{D}_1 \downarrow$ | | RMSE on $\mathcal{D}_2 \downarrow$ | | SC $\uparrow$ | | CA$\downarrow$ | Precision$\uparrow$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AZ | EL | AZ | EL | AZ | EL | | $\mathrm{MP}_{10}$ | $\mathrm{MP}_{25}$ |
| DML | **Metric L.** | - | - | - | - | 0.85 | 0.49 | 0.16 | 0.88 | 0.88 |
| | **+ SSL** | 9.47 | 8.80 | 18.90 | 7.65 | 0.94 | 0.65 | 0.18 | **0.93** | **0.91** |
| | **+ Bin. dec.** | - | - | - | - | 0.91 | 0.50 | **0.15** | 0.89 | 0.88 |
| | **+ Combined** | **8.30** | **6.90** | **14.45** | **7.15** | **0.96** | **0.66** | **0.15** | **0.93** | 0.90 |
| SSL | **Classif.** | 11.90 | 9.20 | 24.20 | 9.41 | 0.93 | 0.51 | 0.24 | 0.92 | 0.88 |
| | **Regres.** | 13.20 | 12.90 | 23.01 | 14.11 | 0.94 | 0.58 | 0.23 | 0.91 | 0.89 |
| Baseline | **BAMQ** | - | - | - | - | 0.16 | 0.02 | 0.79 | - | - |
| | **DPLM** | 19.45 | 23.54 | 18.99 | 24.56 | 0.93 | 0.25 | 0.25 | 0.87 | 0.79 |
| | **SAQAM** | 17.01 | 19.26 | 17.29 | 20.21 | 0.94 | 0.41 | 0.19 | 0.92 | 0.89 |
| | **GPSM+BMFD** | - | - | - | - | 0.93 | 0.56 | 0.17 | - | - |

Table 1: ***Objective evaluations**: Models evaluated include: DML models, SSL models (e.g., classification and regression), and baseline models including BAMQ, DPLM, SAQAM, and GPSM+BMFD. $\uparrow$ or $\downarrow$ is better.*

| Type | Name | P1 | | | P1' | | P2 | P3 | | | P4 | | | | | | 2AFC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Speech | Castanets | Guitar | Speech | Castanets | Music | Speech | Pink Noise | Guitar | Pink Noise | Vocals | Castanets | Glocken | EM | AM | |
| DML | **Original** | 0.84 | 0.83 | 0.87 | 0.57 | 0.78 | 0.48 | 0.55 | 0.19 | 0.17 | 0.47 | 0.49 | 0.45 | 0.45 | 0.49 | 0.69 | 82.80 |
| | **+SSL** | 0.92 | 0.88 | 0.94 | 0.81 | 0.94 | 0.71 | **0.81** | 0.79 | 0.49 | **0.58** | 0.56 | 0.50 | 0.48 | 0.61 | 0.79 | 83.20 |
| | **+Bin. dec** | 0.89 | 0.85 | 0.91 | 0.75 | 0.89 | 0.65 | 0.64 | 0.45 | 0.31 | 0.49 | 0.51 | 0.45 | 0.46 | 0.57 | 0.76 | 85.01 |
| | **+Combined** | **0.96** | **0.97** | **0.97** | **0.88** | 0.99 | 0.79 | **0.81** | 0.79 | **0.61** | **0.58** | **0.65** | **0.51** | **0.56** | **0.69** | **0.89** | **87.95** |
| SSL | **Classif.** | **0.96** | 0.95 | 0.89 | 0.85 | 0.97 | 0.48 | 0.75 | 0.25 | 0.18 | 0.50 | 0.59 | 0.39 | 0.47 | 0.49 | 0.81 | 77.20 |
| | **Regres.** | 0.94 | 0.96 | 0.93 | 0.87 | 0.94 | 0.51 | 0.79 | 0.21 | 0.21 | 0.43 | 0.60 | 0.39 | 0.48 | 0.51 | 0.81 | 80.18 |
| Baseline | **BAMQ** | 0.03 | 0.83 | 0.09 | 0.52 | 0.77 | -0.17 | 0.42 | 0.65 | 0.08 | -0.02 | 0.36 | 0.11 | -0.05 | 0.23 | 0.18 | 60.75 |
| | **DPLM** | 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 | 78.96 |
| | **SAQAM** | **0.96** | 0.95 | 0.94 | **0.88** | **1.0** | 0.52 | 0.78 | 0.23 | 0.26 | 0.53 | 0.60 | 0.49 | 0.47 | **0.69** | 0.82 | 79.12 |
| | **GPSM+BMFD** | 0.94 | 0.96 | 0.89 | 0.35 | 0.86 | **0.83** | 0.71 | **0.94** | 0.58 | **0.58** | 0.60 | 0.41 | 0.53 | 0.57 | 0.83 | 83.20 |

Table 2: ***Subjective evaluation****: Models evaluated include DML models, SSL models (e.g., classification and regression), and baseline models including BAMQ, DPLM, SAQAM, and GPSM+BMFD. Spearman Correlation (SC).* $\uparrow$ *is better.*

We select four distinct classes of datasets: (1) Bilateral Ambisonics (P1 and P1') from [32]; (2) Spherical Microphone Array (P2) from [33]; (3) Headphone Equalization (P3) from [5]; and (4) Bitrate Compressed Ambisonics (P4) from [6]. We compute Spearman correlation (SC) score between the model's predicted distance with the publicly available subjective ratings. The results is shown in Table 2. Note that, in addition to evaluating on speech recordings, we included experiments on generic audio signals (Castanets, Guitar, Music, Pink Noise, etc.) to show that our model generalizes across domains without any specific fine-tuning for any objective or subjective evaluation task.

Additionally, we conduct a 2-Alternative Forced Choice (2AFC) test where we presented one reference and two test recordings and asked listeners which one of the test recording sounded more similar to the reference. Roughly five listeners evaluate each triplet. We calculate our models' output on each of these triplets and report a ratio of how many of these it follows (shown in Table 2). Overall, we see that our DML models improve in correlation with subjective ratings as more tasks are added. In most of the objective and subjective evaluations, the DML model performs better than SSL-based approaches. This is primarily due to the fact that DML approaches are trained for content in-variance explicitly, whereas SSL-based approaches are not. Since DML-based approaches learn a better separation between content and spatial cues, they tend to generalize better. Moreover, we also see the regression-based SSL model performs better than the classification-based SSL model suggesting that optimizing for finer estimates leads to a better-performing model. It is worth noting that the best-performing model overall is the DML model with both SSL and Binaural-decoder tasks combined, which we call the **SQM-BS** metric.

### 4.3. Ablation analysis

**2-channel inputs** In most cases, we may have the mono channel present corresponding to the binaural signal. To investigate the usefulness of the 3-channel (binaural + mono) input, we train another model solely on binaural channels as inputs. Table 3 shows the results on localization error and correlation with subjective ratings. We see that the 2-channel input model performs inferior to the 3-channel model, especially in elevation localization. This is primarily because our model is designed to extract inter-channel information between the L and mono, R and mono, and L and R channels, thereby enforcing the inter-aural time and phase differences between the inputs and disambiguating between source content and loudness level changes with depth changes. For example, if the mono signal is amplified directly by 10dB, the signal received at the left and right ears will be amplified relatively by the same amount. Because our model architecture learns the relative differences (for example, between mono-L, mono-R, and L-R), there is no relative gain.

| Name | RMSE $\mathcal{D}_1 \downarrow$ | | P1 $\uparrow$ | | | P2 $\uparrow$ | P3 $\uparrow$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | AZ | EL | Speech | Castanets | Guitar | Music | Speech | Pink Noise | Guitar |
| **SQM-BS** | **8.30** | **6.90** | **0.96** | **0.97** | **0.97** | **0.79** | **0.81** | **0.79** | **0.61** |
| **2ch input** | 14.45 | 16.21 | 0.57 | 0.69 | 0.75 | 0.45 | 0.42 | 0.45 | 0.14 |
| **Low pass** | | | | | | | | | |
| **+ 16k** | 10.90 | 13.90 | 0.83 | 0.92 | 0.94 | 0.70 | 0.76 | 0.59 | 0.36 |
| **+ 8k** | 11.40 | 14.90 | 0.77 | 0.90 | 0.91 | 0.66 | 0.71 | 0.51 | 0.32 |

Table 3: ***Ablation studies****. Sec 4.3 describes RMSE* $\mathcal{D}_1$*, and Spearman correlations across 3 datasets.* $\uparrow$ *or* $\downarrow$ *is better.*

**Sampling rate**: In order to assess the performance gains using a model trained on higher fidelity inputs (e.g., full bandwidth signals with 48kHz), we train three models on data that have been down-sampled to 8kHz and 16kHz. We observe a trend that suggests that the performance (localization error, MOS correlation, etc.) increases as the sampling frequency increases (see Table 3).

| | PESQ$\uparrow$ | STOI$\uparrow$ | L2$\downarrow$ | M.STFT$\downarrow$ | Si-SDR$\uparrow$ |
|---|---|---|---|---|---|
| **Noisy** | 1.15 | 70.90 | 0.058 | 0.32 | -1.51 |
| **L1** | 1.65 | 83.60 | 0.013 | 0.18 | 9.15 |
| **SQM-BS** | **2.19** | **88.50** | **0.011** | **0.10** | **11.20** |

Table 4: **Evaluation of enhancement** *models using a held-out test set with objective measures.*

### 4.4. Binaural speech enhancement

To demonstrate the utility of our metric as a differentiable loss, we took the binaural speech enhancement (SE) model designed in [20] as a baseline. We applied our best-performing SQM-BS model as a loss function to train the SE system from scratch. We followed the same training routine described in the original work. We evaluated the final SE model on unseen audio recordings from a dataset [34, 35]. Following prior works in SE, we evaluate the quality of enhanced binaural recordings using a variety of objective measures: i) PESQ; (ii) Short Time Objective Intelligibility (STOI); (iii) L2 distance on the waveform; (iv) Scale-invariant signal to distortion ratio (Si-SDR); and v) Multi-resolution STFT evaluated over each channel separately and then averaged to get a single value. As shown in Table 4, SE model trained with SQM-BSas a loss function have higher scores than the baseline. This highlights the usefulness of our proposed set of models in audio similarity tasks, especially in identifying and eliminating minor human perceptible artifacts that are not captured by traditional losses.

## 5. Conclusions and Future work

We presented a new metric SQM-BS that assesses localization similarity between binaural signals based on triplet learning and multi-task learning using sound source localization (SSL) and binaural speech synthesis as two additional tasks. We compare our proposed approach to existing baseline models in terms of several properties such as localization errors in (un)seen environments, monotonicity with increasing angular distance, invariance to content, and retrieval performance, and find that our proposed approach outperforms them. We also demonstrated the utility of our metric as a differentiable loss function. Because our model was trained on speech signals, its application is limited to speech recordings; however, future research could extend our model to handle other types of sounds, such as music.

# 6. References


[1] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *ITU Assembly*, 2014.

[2] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality—technology and applications," *IEEE TASLP*, vol. 14, no. 6, pp. 1890–1901, 2006.

[3] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM TASLP*, vol. 29, pp. 1530–1541, 2021.

[4] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acustica*, vol. 100, no. 5, pp. 984–994, 2014.

[5] I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The effect of generic headphone compensation on binaural renderings," in *AES ICII*, 2019.

[6] T. Rudzki, I. Gomez-Lanzaco, P. Hening, J. Skoglund, T. McKenzie, J. Stubbs, D. Murphy, and G. Kearney, "Perceptual evaluation of bitrate compressed ambisonic scenes in loudspeaker based reproduction," in *AES ICII*, 2019.

[7] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2607–2624, 1992.

[8] E. H. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596.

[9] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of ASA*, vol. 136, no. 2, pp. 791–802.

[10] R. Huber and B. Kollmeier, "PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception," *IEEE TASLP*, vol. 14, no. 6, pp. 1902–1911, 2006.

[11] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the AES*, no. 6, 2013.

[12] S. Kampf, J. Liebetrau, S. Schneider, and T. Sporer, "Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio," *Journal of the AES*, october 2010.

[13] J.-H. Fleßner, R. Huber, and S. D. Ewert, "Assessment and prediction of binaural aspects of audio quality," *Journal of AES*, vol. 65, no. 11, pp. 929–942, 2017.

[14] J.-H. Seo, S. B. Chon, K.-M. Sung, and I. Choi, "Perceptual objective quality evaluation method for high quality multichannel audio codecs," *Journal of AES*, vol. 61, no. 7/8, pp. 535–545, 2013.

[15] 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.

[16] M. Schäfer, M. Bahram, and P. Vary, "An extension of the PEAQ measure by a binaural hearing model," in *ICASSP*, 2013.

[17] J. Blauert and S. Hearing, "The psychophysics of human sound localization," in *Spatial Hearing*. MIT Press, 1997.

[18] P. Manocha, A. Kumar, B. Xu, A. Menon, I. D. Gebru, V. K. Ithapu, and P. Calamia, "DPLM: A deep perceptual spatial-audio localization metric," in *2021 IEEE WASPAA*. IEEE, 2021, pp. 6–10.

[19] T. Biberger and S. D. Ewert, "Towards a generalized monaural and binaural auditory model for psychoacoustics and speech intelligibility," *arXiv preprint arXiv:2106.15659*, 2021.

[20] P. Manocha, A. Kumar, B. Xu, A. Menon, I. D. Gebru, V. K. Ithapu, and P. Calamia, "SAQAM: Spatial audio quality assessment metric," Interspeech, 2022.

[21] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[22] I. D. Gebru, D. Marković, A. Richard, S. Krenn, G. A. Butler, F. De la Torre, and Y. Sheikh, "Implicit hrtf modeling using temporal convolutional networks," in *ICASSP 2021-2021 IEEE*. IEEE, 2021, pp. 3385–3389.

[23] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*. Springer, 2015, pp. 84–92.

[24] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM TASLP*, vol. 23, no. 7, pp. 1172–1183, 2015.

[25] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," *Interspeech*, 2020.

[26] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[27] B. I. Bacila and H. Lee, "360° binaural room impulse response (BRIR) database for 6DOF spatial perception research," in *AES Convention 146*. AES, 2019.

[28] C. Mittag, M. Böhme, and S. Werner, "Dataset of KEMAR-BRIRs measured at several positions and head orientations in a real room," Dec.

[29] J. a. Francombe, "IoSR listening room multichannel BRIR dataset."

[30] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[31] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 351–355.

[32] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs," *IEEE/ACM TASLP*, vol. 29, pp. 901–913, 2021.

[33] T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. Ahrens, "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, pp. 428–440, 2020.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[35] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC hrtf database," in *Proceedings of the 2001 IEEE WASPAA (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.