OPEN ACCESS Freely available online

CC BY

D. Shariat Panah, D. Barry, A. Ragano, J. Skoglund, and A. Hines,
"BINAQUAL: A Full-Reference Objective Localization Similarity Metric for Binaural Audio,"
J. Audio Eng. Soc., vol. 74, no. 1/2, pp. 20–34 (2026 Jan./Feb.).
https://doi.org/10.17743/jaes.2022.0243.

BINAQUAL: A Full-Reference Objective Localization Similarity Metric for Binaural Audio

DAVOUD SHARIAT PANAH,[1,*] (davoud.shariatpanah@ucd.ie) **DAN BARRY,**[1] (danbarry@duck.com) **ALESSANDRO RAGANO,**[1] (alessandro.ragano@ucd.ie)

JAN SKOGLUND,[2] *AES Member* (jks@google.com) **AND ANDREW HINES,**[1] *AES Member* (andrew.hines@ucd.ie)

[1]*School of Computer Science, University College Dublin, Dublin, Ireland*
[2]*Google LLC, San Francisco, CA*

Spatial audio enhances immersion in applications such as virtual reality, augmented reality, gaming, and cinema by creating a three-dimensional auditory experience. Ensuring the spatial fidelity of binaural audio is crucial, given that processes such as compression, encoding, or transmission can alter localization cues. While subjective tests are used for evaluating spatial localization quality, they are costly and time-consuming. This paper introduces BINAQUAL, a full-reference objective metric designed to assess localization similarity in binaural audio recordings. BINAQUAL adapts the AMBIQUAL metric, originally developed for localization quality assessment in ambisonics audio format, to the binaural domain. BINAQUAL is evaluated across five key research questions, examining its sensitivity to variations in sound source locations, angle interpolations, surround loudspeaker layouts, audio degradations, and content diversity. Results demonstrate that BINAQUAL effectively differentiates between subtle spatial variations and correlates strongly with subjective listening tests, making it a reliable metric for binaural localization quality assessment. The proposed metric provides a robust benchmark for ensuring spatial accuracy in binaural audio processing, paving the way for improved objective evaluations in immersive audio applications.

Keywords: spatial audio, binaural audio, localization quality metric, audio quality assessment

1 INTRODUCTION

Spatial audio is an advanced auditory technology that creates a three-dimensional sound field, enabling listeners to perceive sounds as originating from specific locations in space. This technology enhances user immersion across applications such as virtual reality (VR), augmented reality (AR), gaming, and cinematic soundscapes. Binaural audio, a form of spatial audio, utilizes two channels to simulate the physical and perceptual cues involved in sound localization. With smartphones, headphones, and wearable devices emerging as dominant platforms for music, movies, and interactive media, there is a growing interest in immersive auditory experiences that can be delivered in a convenient and effective manner [1]. In this context, binaural audio is expected to see significant growth, driven by evolving media consumption patterns.

The rise of open-source spatial audio formats, such as Eclipsa Audio [2], alongside binaural remixing of content for headphone playback, is reshaping the way immersive sound is created and experienced. Such innovations highlight the growing need for robust localization quality metrics to assess and optimize spatial accuracy in binaural audio rendering. As content creators and engineers increasingly adopt object-based audio mixing and dynamic head-tracked rendering, existing metrics often fail to reflect perceptual accuracy across diverse playback environments. Developing quantitative and perceptually aligned localization quality metrics is crucial for ensuring consistent auditory immersion in VR, gaming, cinema, and streaming services, where precise spatial positioning enhances realism and listener engagement.

Localization quality assessment refers to the process of quantifying the accuracy and perceptual fidelity of spatial audio reproduction, typically by comparing perceived sound source locations with their intended positions. While Multiple Stimuli with Hidden Reference and Anchor

*To whom correspondence should be addressed, email: davoud.shariatpanah@ucd.ie.

(MUSHRA)–style protocols [3] are occasionally adapted for evaluating spatial attributes, localization accuracy is more commonly assessed using task-specific methods that directly measure directional perception, such as directional pointing tasks (e.g., Makous and Middlebrooks [4]) or minimum audible angle thresholds [5]. However, such tests are time-consuming, expensive, and prone to variability due to individual listener differences and environmental factors [6]. As the demand for reliable and reproducible evaluations grows, effective objective metrics are needed to assess spatial localization quality in diverse contexts.

Developing effective localization quality metrics is a challenging task. Factors such as variabilities in individual listeners, recording environments, and stimulus characteristics can significantly influence the spatial cues in binaural recordings [7]. In recent years, a variety of spatial localization quality metrics have been proposed. Such metrics can be broadly categorized into signal processing–based techniques [8–11], and data-driven models [12–15]. However, the robustness of the metrics to variations in binaural recordings remains poorly understood, primarily due to the lack of rigorous benchmarks. Also, the reproducibility of these metrics is limited, as their source codes are often unpublished.

To address these gaps, this paper introduces a full-reference objective metric called BINAQUAL, designed to assess spatial localization similarity (LS) in binaural audio recordings. Given a binaural audio signal and its corresponding reference, BINAQUAL provides a score that measures how perceptually similar their spatial localization characteristics are. BINAQUAL is developed by adapting an objective localization quality metric originally designed for ambisonic audio signals [16]. Through extensive experiments and comprehensive benchmarking, the robustness of the proposed metric is investigated with respect to environmental and stimulus-related factors. To achieve this, multiple synthetic datasets tailored to the research objectives of the study are generated. As a key contribution of this work, a combined dataset, along with the source code of the metric, is compiled and publicly released to facilitate further research and ensure reproducibility.

To rigorously evaluate the BINAQUAL, its sensitivity to small variations in the spatial locations of point sources is first examined. If the metric fails to distinguish between sounds originating from different locations, it may instead be responding to differences in fidelity, intensity, or other confounding factors rather than localization cues. Understanding this sensitivity is fundamental to ensuring the metric's validity for spatial localization evaluation.

Second, whether the metric can differentiate between binaural renders that use real versus interpolated impulse responses is examined. Ideally, binaural rendering systems would have access to real impulse responses for all source angles, but in practice, interpolation is often used to approximate missing data. Evaluating the ability of the model to distinguish between real angle versus angle interpolation renders ensures that it is capable of assessing the accuracy of binaural renderings.

Third, whether the metric can distinguish between simulated binaural representations of different surround loudspeaker layouts is investigated. Many spatial audio recordings are mixed specifically for distinct surround layouts, such as 5.1 or 7.1. When down-mixed for binaural listening, preserving spatial cues is essential. A metric capable of distinguishing between these layouts ensures that the spatial localization of the audio is accurately preserved during the binaural rendering of multichannel surround-encoded audio.

Fourth, the metric's ability to distinguish between different levels of codec compression applied to binaural audio recordings is evaluated. Compression algorithms can introduce distortions that impact spatial cues; hence, a reliable localization quality metric should be able to capture these degradations.

Real-world audio scenes often contain complex spatial distributions of multiple sources, making it essential for the metric to perform reliably under diverse conditions. Therefore, the robustness of the metric is assessed across different types of content and varying numbers of point sources to ensure its reliability and practicality.

To address the research objectives discussed above, the following research questions (RQs) are posed:

- RQ1: Is the proposed metric sensitive to variations in the spatial locations of the point sources?
- RQ2: Can the metric differentiate between single-point source binaural renders that use real versus interpolated angle impulse responses?
- RQ3: Can the metric distinguish between the simulated binaural representations of different surround loudspeaker layouts?
- RQ4: Can the metric distinguish between different codec compression levels of binaural audio recordings?
- RQ5: How robust is the metric to variations in content and number of point sources in binaural audio recordings?

While several localization quality metrics have been proposed in previous work, they primarily focus on accuracy prediction without addressing factors affecting localization robustness. To advance spatial audio quality research beyond simple accuracy estimation, key RQs that have not been explicitly addressed in prior work are investigated. Addressing these questions allows for establishing the robustness, reliability, and practical utility of the BINAQUAL across diverse scenarios. A metric that can reliably evaluate localization quality under these conditions will be more applicable across a broad range of use cases, ensuring that it is practical for real-world deployment. By addressing the above RQs, the BINAQUAL metric sets a benchmark for assessing localization quality in binaural audio with accuracy, reliability, and versatility.

The remainder of this paper is organized as follows. SEC. 2 provides the background on binaural audio and sound source localization, along with a review of existing spatial localization quality metrics. Additionally, this

section includes a summary of the AMBIQUAL metric, which has been adapted in this study. SEC. 3 introduces the BINAQUAL model for assessing spatial LS in binaural audio signals. SEC. 4 describes the evaluation datasets used in the study. In SEC. 5, the results are presented and discussed. Finally, SEC. 6 concludes the paper and outlines potential future directions.

2 BACKGROUND

2.1 Binaural Audio

Binaural audio is an audio format designed to replicate the way humans naturally perceive sound, creating a realistic and immersive auditory experience. Unlike stereo audio, which typically focuses on directional cues in a flat plane, binaural audio simulates the three-dimensional sound field by incorporating spatial cues [17]. This allows listeners to perceive sounds as originating from specific locations in space as if they were physically present in the sound environment.

The perceived authenticity of binaural audio relies on the accurate simulation of spatial auditory cues that humans use to localize sound sources. Among these, interaural time differences (ITD) and interaural level differences (ILD) play key roles in lateral localization. ITD refers to the small delay in the arrival time of a sound at one ear compared to the other and is most effective for low-frequency sounds. ILD describes the difference in sound intensity between the ears, primarily due to the head shadowing effect, and is more prominent for high-frequency sounds [18, 19]. However, these are not the only cues involved in spatial perception. For a plausible and perceptually authentic binaural experience, additional factors such as spectral shaping by the outer ear and dynamic cues related to head movement also contribute significantly to sound source localization in elevation and distance [20, 21]

Binaural audio can be created through recording or synthesis. Binaural recording involves capturing audio using specially designed microphones placed inside the ear canals of a real or artificial head. However, binaural recordings are inherently static; once recorded, the spatial cues are fixed, preventing listeners from interacting with the soundscape by moving their heads or changing their positions [22]. Additionally, the characteristics of the head used in the recording process may not match those of individual listeners, leading to inaccuracies in spatial perception [17]. As a result, this technique has limited applicability in dynamic or personalized real-world scenarios.

In the binaural synthesis technique, spatial audio is synthesized by processing sound signals through head-related transfer functions (HRTFs). The HRTF describes how an individual's body, particularly the head, torso, and outer ears filter sound as it travels from a source to the eardrum [17]. By applying the appropriate HRTFs based on the intended location of each sound source, it is possible to create a binaural representation of a complex sound field. Binaural synthesis allows for dynamic and customizable spatial audio experiences but depends on the accuracy and availability of high-quality HRTF datasets [22]. The experiments in this study use the SADIE II database [23] to synthesize binaural audio samples. This database offers a comprehensive collection of HRTFs that cover a wide range of azimuths and elevations, enabling accurate spatial representation of sound sources.

2.2 Sound Source Localization

Sound source localization is the process of identifying the spatial location of a sound source, including its azimuth, elevation, and distance, based on auditory cues extracted from the sound waves received by a hearing system or sensor array [21]. This process leverages differences in time, intensity, and frequency content between signals captured at multiple points, such as ears or microphones, to determine the source's location in a three-dimensional space.

Multiple factors influence the accuracy of sound source localization in humans. The first factor is individual listener characteristics such as age, head and ear anatomy, and prior auditory experience. It has been shown that age negatively impacts the accuracy of sound source localization in older adults in terms of both azimuth and elevation [24]. Individual differences in head size, ear shape, and torso affect HRTFs, which describe how sounds are filtered by the body before reaching the eardrum. These variations influence localization accuracy [25, 26]. In addition to these, prior auditory experience, such as familiarity with a sound or training in spatial hearing, enhances the ability to interpret localization cues [27].

Characteristics of the environment are another factor that influences localization accuracy in humans. Reverberation occurs when sound waves reflect off surfaces, creating multiple delayed copies of the original sound. Reverberant environments blur the ITD, ILD, and spectral cues, which result in larger errors in sound source localization [28]. The presence of background noise in the environment can also reduce localization accuracy by masking the target sound and reducing the ability to extract spatial cues [29, 30].

Lastly, the characteristics of the stimulus also significantly influence the localization accuracy. Makous and Middlebrooks [4] investigated the influence of sound location on localization error in humans. They found that sound localization is very accurate for sources located directly in front of the listeners, where the average error was 2.0° and 3.5° for horizontal and vertical dimensions, respectively. However, as the sound sources were shifted toward peripheral locations, the errors increased significantly, reaching up to 20° at the extreme positions.

The impact of the spectral content of the stimulus on the accuracy of sound source localization has also been investigated in several studies. It has been shown that both frequency content and bandwidth of the signal affect the accuracy of localization [31–33]. High-frequency spectral cues improve elevation judgments and help resolve front-back confusion [34, 35]. Localization accuracy for low-frequency sounds is typically good in azimuth but poorer in elevation due to the lack of spectral cues [34, 36]. Regarding the sound bandwidth impact, previous work showed

that wider stimulus bandwidth leads to higher localization accuracy and vice versa [32, 33, 37]. This is because signals encompassing a wider range of frequencies provide richer spatial cues, leading to improved localization accuracy.

A wide range of sound source localization models have been proposed [38–41]. Some of these models, such as those by Lindemann [42] and Baumgartner et al. [43], are available in the MATLAB Auditory Modeling Toolbox.

2.3 Spatial Localization Quality Metrics

Significant efforts have been made to develop objective metrics for assessing localization quality in spatial audio content. The work in this domain can be broadly categorized into signal processing–based and data-driven techniques.

Choi et al. [8] proposed a model that leverages ILD, ITD, and interaural cross-correlation coefficient distortions to measure perceived quality in multichannel audio. Fleßner et al. [9] introduced BAM-Q, a metric that uses a binaural auditory model as a front-end to extract ILD, ITD, and interaural coherence from binaural signals, which are then utilized for spatial quality assessment. Joint modeling of monaural and binaural signal distortions has also been explored. For example, the work in [10] combines BAM-Q with a generalized power-spectrum model to predict quality based on both types of distortions. More recently, Watcharasupat and Lerch [11] presented a spatial quality evaluation model using a filter decomposition technique based on the duplex theory of spatial hearing. In another recent work, Ren et al. [44] proposed an objective metric called KSRIR to predict the listening quality and localization accuracy of first-order ambisonics generated using interpolated room impulse responses. The quality scores were computed by comparing the cumulative distribution functions of the reference and synthesized signals using the Kolmogorov-Smirnov test.

An early contribution to data-driven approaches was made by Fleßner et al. [12], who proposed a spatial quality assessment model based on multivariate adaptive regression splines. Manocha et al. [13] introduced DPLM, a full-reference spatial localization quality metric based on a direction-of-arrival deep network model. They extended this work with SAQAM [14], a multitask learning model that evaluates both spatial localization quality and listening quality of binaural signals. However, these models were limited to audio signals with a 16-kHz sampling rate. To address this limitation, the authors recently proposed a new spatial quality assessment metric for binaural speech [15]. This metric leveraged a multitask learning framework with direction-of-arrival estimation and binaural speech synthesis as auxiliary tasks. While this model is sampling-rate agnostic, its source code is not available, which limits the reproducibility of the work. More recently, Zheng et al. [45] proposed a spatial quality metric called HAPG-SAQAM for evaluating the spatial and overall audio quality of binaural signals. Their method uses gammatone frequency cepstral coefficients along with perceptual weighted loss, showing improvements over the SAQAM metric.

2.4 AMBIQUAL

AMBIQUAL [16] is a full-reference objective metric designed to assess the quality of compressed ambisonic spatial audio. It evaluates both listening quality and localization accuracy directly from B-format ambisonic. More specifically, listening quality is determined from the omnidirectional channel, while localization accuracy is calculated based on a weighted product of similarities from the directional channels in the B-format audio.

AMBIQUAL was developed by adapting and extending ViSQOL [46], a full-reference objective metric designed to evaluate the perceptual quality of audio signals. ViSQOL operates by comparing a reference (high-quality) audio signal with a test (degraded) version. For this purpose, it uses the Neurogram Similarity Index Measure (NSIM) [47], which measures the similarity between corresponding time-frequency components of the reference and test signals, providing a quantitative indication of how well the test signal preserves the spectral characteristics of the reference.

Unlike ViSQOL, which uses magnitude spectrograms to measure similarities between reference and test signals, AMBIQUAL leverages spectrograms of phase angles (phaseograms). Phaseograms are segmented into fixed-duration patches, and NSIM is applied locally to each aligned reference and test patch. The final similarity score is obtained by averaging NSIM values across all patches and frequency bands. These scores are aggregated across directional channels to estimate localization fidelity.

AMBIQUAL quality predictions were compared with the results of a MUSHRA subjective listening test for a varied set of compressed ambisonics contents, and it was shown that AMBIQUAL predictions correlate well with scores from the listening test [16]. However, the AMBIQUAL model is limited to ambisonic content, as it operates on B-format rather than the binaural render. In other words, the model provides a single quality score for an ambisonic recording, regardless of the render angle.

3 BINAQUAL MODEL

BINAQUAL is a full-reference objective metric for assessing spatial LS in binaural audio recordings. BINAQUAL is designed by adapting the AMBIQUAL model to two-channel binaural audio signals. In practice, BINAQUAL could be applied to an ambisonic signal but only for a particular render. The hypothesis was that adapting the AMBIQUAL model and applying it at the post-render stage could allow any binaurally rendered spatial audio signals to be evaluated, regardless of their origin mix or format.

As mentioned earlier, in AMBIQUAL, localization accuracy is calculated using a weighted product of similarities from directional channels of B-format ambisonics. Given that binaural audio signals consist of only two channels (left and right), BINAQUAL calculates LS as the product of similarities between the reference and test signals in these two channels. It should be noted that BINAQUAL is a relative LS measure, not an absolute localization error estimator (i.e., angular degrees).

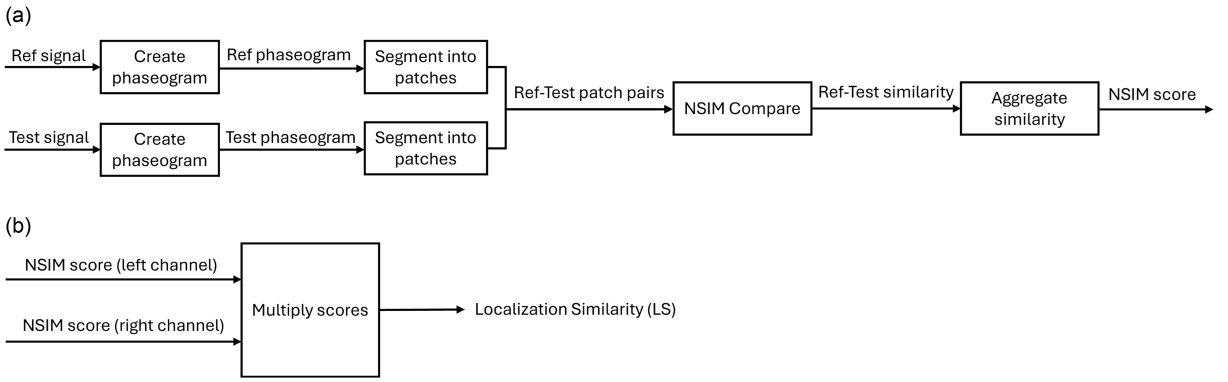


Fig. 1. A high-level representation of the BINAQUAL model. (a) The process to calculate NSIM scores for each channel of reference and test signals. (b) The process to calculate LS from NSIM scores.

Fig. 1 illustrates a high-level representation of the BINAQUAL model. As shown in Fig. 1(a), the process for calculating the NSIM score for reference and test signals resembles the approach used in the AMBIQUAL model. A summary of this process is provided below, and readers are encouraged to refer to [16] for additional details.

As shown in Fig. 1(a), spectrograms of phase angles (phaseograms) are created for each channel of the reference and test signals. These phaseograms are computed using a 2,048-point short-time Fourier transform (STFT) with a 1,536-point Hamming window and 50% overlap. To align the analysis with the perceptually motivated frequency range of the 32-band gammatone filterbank (50 Hz to 16 kHz), a specific subset of STFT frequency bins is selected. The 640 bins corresponding to indices 2–641 of the original STFT are used. At a 48-kHz sampling rate, this selection captures a frequency range from approximately 47 Hz to 15.02 kHz. This range effectively covers the bands of interest from the gammatone filterbank while excluding the DC component (bin 0).

NSIM is then used to compute localized similarity between patch pairs using two components: an intensity term that reflects the similarity of local means, and a structure term that captures the correlation (covariance) between patches. Both components are stabilized with constants derived from the dynamic range, which, in the case of phaseograms, corresponds to 2π. These frame-wise similarity scores are computed over all 640 frequency bins and then averaged within 32 perceptually motivated gammatone frequency bands. The band-averaged scores are further aggregated (averaged) across time patches to produce a single global similarity score per channel. The final NSIM score ranges from 0 to 1, where 1 indicates identical patches and values closer to 0 indicate low similarity. This bounded nature makes NSIM particularly well-suited for perceptual similarity tasks, offering normalized and interpretable scores, unlike unbounded metrics such as root mean squared error. Readers are encouraged to refer to [47] for further details about NSIM.

This process is repeated for both channels of the reference and test signals. As shown in Fig. 1(b), the LS score is calculated as the product of the final NSIM scores from the two channels.

The motivation for applying NSIM to phaseograms of binaural audio signals lies in the relationship between phase information and the spatial cues essential for sound localization. As mentioned earlier, ITDs dominate low-frequency localization, while ILDs dominate at high frequencies. These cues are encoded in the temporal and spectral structure of binaural signals. By comparing the phaseograms of left and right channels between reference and test signals, BINAQUAL indirectly assesses the fidelity of the spatial cues, particularly ITD and ILD, embedded in those signals. Since the NSIM metric is sensitive to both local phase distortions and structural mismatches across frequency bands, it provides a perceptually relevant measure of how well the test signal preserves spatial localization characteristics. Thus, high similarity between binaural phaseograms implies preserved localization cues, whereas reduced similarity suggests perceptual degradation in spatial fidelity. The source code of the BINAQUAL model is available online.[1]

4 EVALUATION DATASETS

To present a comprehensive evaluation of the BINAUQAL model, three evaluation datasets were created. These datasets were generated synthetically by applying specific HRTFs from subject D2 (mannequin head) of the SADIE II database [23] to various audio contents using the Binamix Python library [48]. When the desired angle was available in the SADIE database, the corresponding binaural impulse response was used directly. For angles not present in the database, suitable binaural impulse responses were generated via interpolation at render time. This was achieved by identifying the nearest available angles using a modified Delaunay triangulation and computing interpolation weights based on Euclidean distances. For further details on the interpolation method, readers are referred to [48].

[1] https://github.com/QxLabIreland/Binaqual.

Table 1. The increments for different azimuth ranges. These values were determined by calculating the average human localization error across different elevations for the following azimuth ranges.

From (degree)	To (degree)	Increment (degree)
0	59	5
60	119	8
120	179	12
180	239	12
240	299	8
300	359	5

Table 2. The increments for different elevation ranges. These values were determined by calculating the average human localization error across different azimuths for the following elevation ranges.

From (degree)	To (degree)	Increment (degree)
0	29	5
30	59	10
60	90	15
0	–29	5
–30	–59	10
–60	–90	15

The datasets are specifically designed to address RQ1–3, as outlined in SEC. 1. The same dataset used in [16] is used to address RQ4. It is worth mentioning that a separate dataset is not created for RQ5; instead, the robustness of the model to various contents and the number of point sources is assessed within the scope of RQ1–4. The compiled dataset is called SynBAD (Synthetic Binaural Audio Dataset) and is available online.[2] A detailed description of each dataset is provided below.

4.1 Localization Sensitivity Dataset

The localization sensitivity dataset is designed to evaluate the model's sensitivity to the locations of point sources (RQ1). To generate the dataset, a subset of HRTFs with varying azimuths and elevations from the SADIE II database was selected. The selection was based on the average human performance in sound source localization. As mentioned in SEC. 2.2, human localization errors tend to increase as sound sources shift toward peripheral locations. To account for this, the average unsigned localization errors reported in [4] across different azimuth and elevation ranges were used to determine a subset of HRTFs to synthesize binaural samples. Tables 1 and 2 summarize the increments applied for different azimuth and elevation ranges, respectively.

To evaluate the robustness of the model to different contents, ten mono audio samples were used to synthesize the dataset. As summarized in Table 3, these audio samples were either synthetically generated or originated from the European Broadcasting Union (EBU) dataset [49] and cover a wide range of frequencies and bandwidths. Three samples with reverberation effect were also included in the base sample set to analyze the model's performance in reverberant conditions.

Using the angles and audio content described above, 14,500 binaural samples were synthesized at a sampling rate of 48 kHz. Since the dataset was intended solely for evaluation and not for subjective listening tests, each sample was limited to a duration of 2 s, sufficient for assessing model sensitivity to point source locations.

4.2 Angle Interpolations Dataset

The angle interpolations dataset is designed to evaluate the model's ability to differentiate between point source renders using discrete angles versus interpolated angles (RQ2). To create this dataset, the audio contents described in Table 3 were binaurally rendered using various degrees of interpolation between angle pairs. These angle pairs have been summarized in Table 4. As shown in Table 4, the first set of angles is used to render a single source virtually at azimuth, elevation = (50°, 0°), while the second set is used for a virtual render at (90°, 0°). As an example, a reference at (50°, 0°) can be virtually represented by interpolation between any of the following angle pairs: {(45°, 0°) and (55°, 0°)}, {(40°, 0°) and (60°, 0°)}, {(30°, 0°) and (70°, 0°)}, {(20°, 0°) and (80°, 0°)}, {(10°, 0°) and (90°, 0°)}, and {(0°, 0°) and (100°, 0°)}. Using the angle pairs described in Table 4, 140 binaural samples with a duration of 5 s and sampling frequency of 48 kHz were synthesized.

4.3 Surround Layouts Dataset

The surround layouts dataset is designed to evaluate the model's ability to distinguish between simulated binaural representations of various surround loudspeaker layouts, including 5.1, 5.1.4, 7.1, and 7.1.4 (RQ3). The previous datasets included sources rendered at a large variety of angles. However, the surround layouts dataset contains renders of various contents at only two sets of angles, one with elevation and one without. This is because this dataset will only be used to investigate the sensitivity of the metric to the presence of elevation (i.e., height loudspeaker) in surround loudspeaker setups.

To generate samples of this dataset, for each of the surround layouts mentioned above, two point sources were rendered at the following angles (azimuth, elevation): (60°, 0°) and (300°, 0°), as well as (60°, 50°) and (300°, 30°). The sources were also rendered at the above angles without any loudspeaker layouts to be used as reference signals.

The Binamix library [48] was used to handle the rendering of sources for specific surround loudspeaker layouts. This library uses impulse response interpolation to simulate the VBAP method [50], rendering the source at the virtual loudspeaker positions between the discrete angles available for a specific loudspeaker layout. The audio contents described in Table 3 were mixed in different ways to ensure a diverse set of sound source combinations in this dataset. These combinations, along with the purpose of each, have been detailed in Table 5. Using the angles and audio contents described above, 80 binaural samples with

[2] https://zenodo.org/records/15431990.

Table 3. Description of the base mono audio samples used for synthesizing the localization sensitivity dataset.

Audio Content	Frequency	Bandwidth	Source
Bursty white noise	Low, Mid, High	Broad	Synthetic
Bursty pink noise	Low, Mid	Broad	Synthetic
Bursty pink noise with reverb	Low, Mid	Broad	Synthetic
Male speech	Low, Mid	Medium	EBU
Female speech	Mid, High	Medium	EBU
Female speech with reverb	Mid, High	Medium	Processed EBU
Castanets	Mid, High	Medium	EBU
Castanets with reverb	Mid, High	Medium	Processed EBU
Glockenspiel	High	Medium	EBU
Pure tone (4 kHz)	High	Narrow	Synthetic

Table 4. The angle pairs used to generate the angle interpolations dataset.

Reference: Azimuth = 50°, Elevation = 0°			
Azimuth 1	Elevation 1	Azimuth 2	Elevation 2
45°	0°	55°	0°
40°	0°	60°	0°
30°	0°	70°	0°
20°	0°	80°	0°
10°	0°	90°	0°
0°	0°	100°	0°
Reference: Azimuth = 90°, Elevation = 0°			
Azimuth 1	Elevation 1	Azimuth 2	Elevation 2
85°	0°	95°	0°
80°	0°	100°	0°
70°	0°	110°	0°
60°	0°	120°	0°
50°	0°	130°	0°
40°	0°	140°	0°

a duration of 5 s and sampling frequency of 48 kHz were synthesized.

4.4 Codec Compression Dataset

The codec compression dataset is employed to test the model's ability to distinguish between different compression levels applied to binaural recordings (RQ4). This dataset contains first-order and third-order ambisonics of various audio contents. The samples include both single-point and two-point audio sources with a variety of localizations, including dynamic and fixed sources. The audio samples were encoded using the Opus 1.2 codec, channel mapping family 2 at various bit rates, including 512, 384, 256, 128, 96, 64, and 32 kbps to produce a range of conditions. They were then rendered binaurally for presentation. Readers are encouraged to refer to [16] for a full description of the dataset samples.

5 RESULTS AND DISCUSSION

This section presents and analyzes the results with respect to five RQs outlined in SEC. 1.

5.1 Localization Sensitivity

LS was computed for each sample in the localization sensitivity dataset, using a reference source localized at azimuth = 0°, elevation = 0°. Fig. 2 illustrates the LS as a function of azimuth and elevation angles for female and castanets sources.

As shown in Fig. 2, for all elevations except ±90°, the LS between test and reference signals decreases as the test azimuth moves further away from the reference. However, an increase in LS is observed between –120° to –180° and 120° to 180°, which can be attributed to front-back confusion. For elevations of ±90°, the LS curve remains relatively flat across azimuths. This is expected, as at ±90° elevation, corresponding to directly above or below the listener, azimuth resolution effectively collapses. In other words, in Cartesian coordinates, all azimuths converge to a single point, making azimuth distinctions meaningless. Additionally, binaural localization cues such as ITD and ILD become minimal at these elevations, and although spectral cues are still present, they offer limited azimuthal discrimination, further contributing to a uniform LS profile.

Fig. 3 shows half-hemisphere heatmaps of LS for all contents in the localization sensitivity dataset. To create these heatmaps, Robinson projection from Cartopy, a Python library for geographic map generation, was used. Since the LS scores were in a small range (mostly between 0 and 0.2), the Box-Cox transform was applied to spread out the distribution, making variations more noticeable. As shown in Fig. 3, LS is highest at the reference point (0°, 0°) and decreases with increasing angular distance. However, LS increases again between azimuths 120° and 180°, indicating front-back confusion. This pattern is consistent across all contents except glockenspiel and pure tone. For the glockenspiel, LS increases between azimuths 60° and 90°. This is likely due to its high-frequency spectral content, which weakens ITD cues. Additionally, the pure tone heatmap shows discrete regions of high LS. This could be because narrow-band, high-frequency pure tones provide weak ITD/ILD cues, leading to inconsistent LS across azimuths.

Overall, the results indicate that the BINAQUAL is sensitive to small variations in the spatial locations of single-point sources. The consistency across different source types also demonstrates the model's robustness to variations in content.

Table 5. Description of the two-point sources used for synthesizing the surround layouts dataset. The last column summarizes the testing purpose of each combination.

Source #1	Source #2	Testing Purpose
Male speech	Female speech with reverb	Localizing two vocal sources with different frequency ranges
Castanets	Male speech	Localizing a percussive natural sound against a vocal source
Bursty pink noise	Female speech	Localizing a synthetic broadband sound against a vocal source
Bursty white noise	Male speech	Localizing a synthetic broadband sound against a vocal source
Bursty pink noise with reverb	Female speech with reverb	Localizing two reverberant sounds with different characteristics
Bursty white noise	Bursty pink noise with reverb	Localizing two synthetic broadband sounds with distinct temporal properties
Bursty pink noise	Castanets	Localizing a synthetic sound against a natural percussive sound
Bursty white noise	Castanets with reverb	Localizing a synthetic sound against a reverberant percussive natural sound

5.2 Angle Interpolations

LS between the samples in the angle interpolations dataset and their corresponding reference signals was calculated. Fig. 4 shows the LS across different angle interpolations for two different scenarios: in Fig. 4(a), reference angle = (50°, 0°) and, in Fig. 4(b), reference angle = (90°, 0°).

As shown in Fig. 4, overall, the LS decreases as interpolated angles deviate from the references. This is expected because larger angular separations introduce greater perceptual localization differences. Bursty white and bursty pink show higher LS scores across all angles. These signals are rich in broadband spectral content, which provides strong localization cues. The pure tone signal has the lowest LS scores, indicating that it is the hardest to localize. This aligns with the known difficulty in localizing pure tones due to phase ambiguities and the lack of strong interaural cues. Glockenspiel also shows lower LS scores for more extreme interpolations. Since it has strong high-frequency content (around 1–8 kHz, with prominent overtones), the model relies more on ILD cues, which are less precise for localization. These results confirm that BIANQUAL can effectively distinguish between single-source simulated representations of different angle interpolations while being robust to variations in content.

5.3 Surround Loudspeaker Layouts

LS between samples in the surround layouts dataset, rendered for various surround loudspeaker layouts, and reference signals (without any loudspeaker layout) was calculated. Fig. 5 presents the aggregated mean LS scores across different loudspeaker layouts for two scenarios: in Fig. 5(a), source angles without elevation and, in Fig. 5(b), source angles with elevation. The mean LS scores were computed across eight different audio samples per layout. For each configuration, the 95% confidence interval was also calculated using the standard error of the mean across the eight samples.

As shown in Fig. 5(a), when the angles have no elevation, the LS scores for 5.1 and 5.1.4 are similar, as are those for 7.1 and 7.1.4. Additionally, LS scores for 7.1 and 7.1.4 are higher than those for 5.1 and 5.1.4. This suggests that increasing the number of surround channels enhances LS due to improved spatial coverage of surround loudspeakers.

As shown in Fig. 5(b), when angles have elevations, the LS scores for layouts with height channel (5.1.4, 7.1.4) are

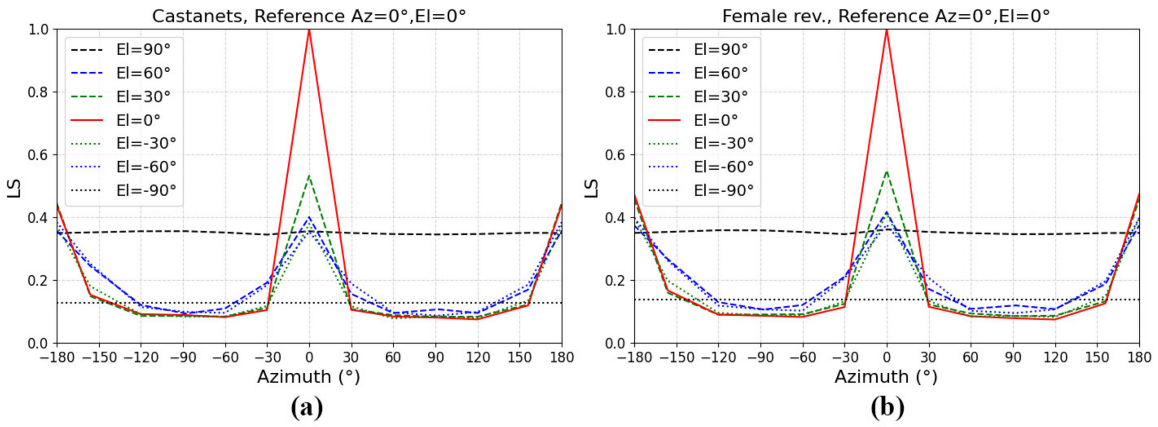


Fig. 2. LS as a function of azimuth and elevation with fixed reference audio source, localized at azimuth = 0°, elevation = 0° for (a) castanets and (b) female speech with reverb test signals.

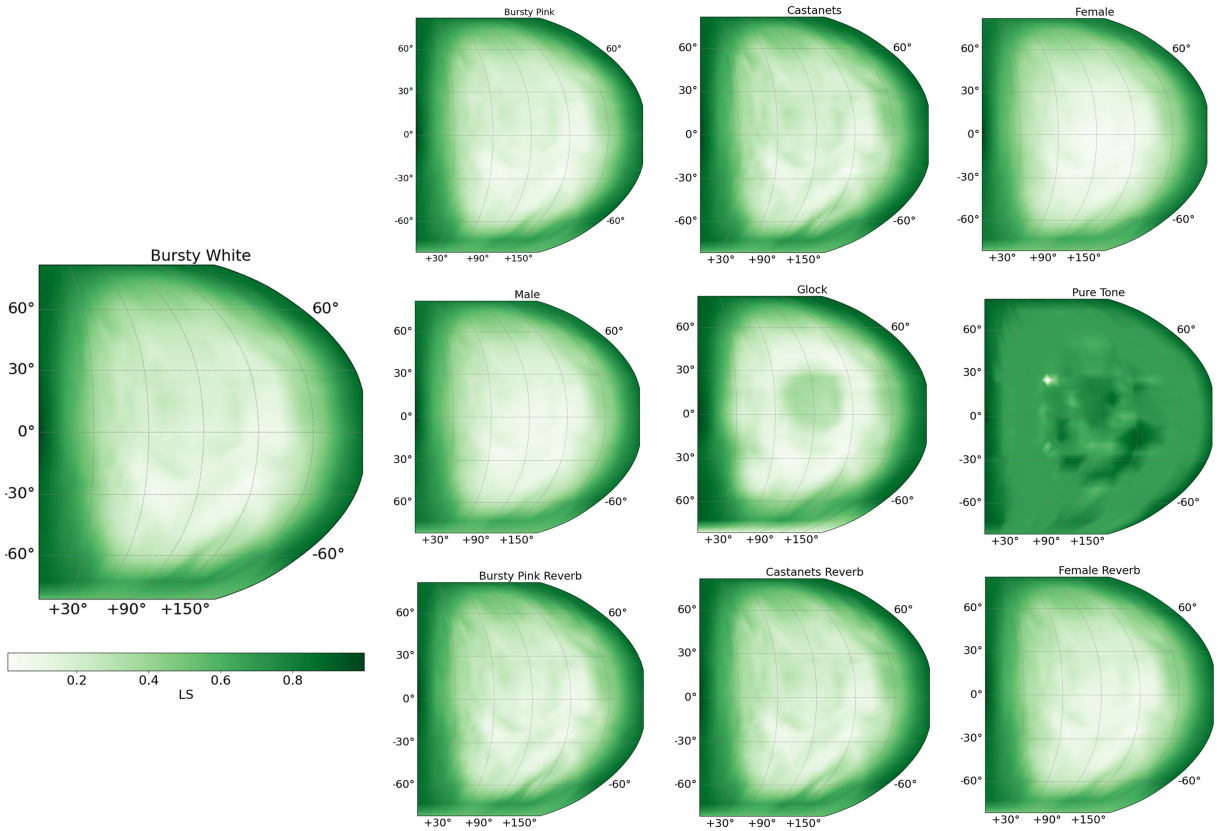


Fig. 3. LS scores distributed on a half sphere for various contents. The corresponding reference audio sources were localized at azimuth = 0°, elevation = 0°. Box-Cox transform was applied to LS scores to spread out the distribution and make variations more noticeable.

higher than those for layouts without elevation (5.1, 7.1). This confirms that elevated angles benefit more from overhead loudspeakers, as adding height channels allows the model to capture vertical localization cues more effectively, leading to an overall increase in LS.

Statistical analysis using nonparametric tests further supports these findings. As shown in Table 6, Mann–Whitney U tests reveal that increasing the number of surround channels leads to significantly higher LS scores, as evidenced by comparisons between 5.1 versus 7.1 and 5.1.4 versus 7.1.4 (both with $p = 0.0002$). Similarly, adding height channels significantly improves LS for both 5.1 versus 5.1.4 ($p = 0.0499$) and 7.1 versus 7.1.4 ($p = 0.0003$). Kruskal–Wallis tests across all layouts further confirm overall sig-

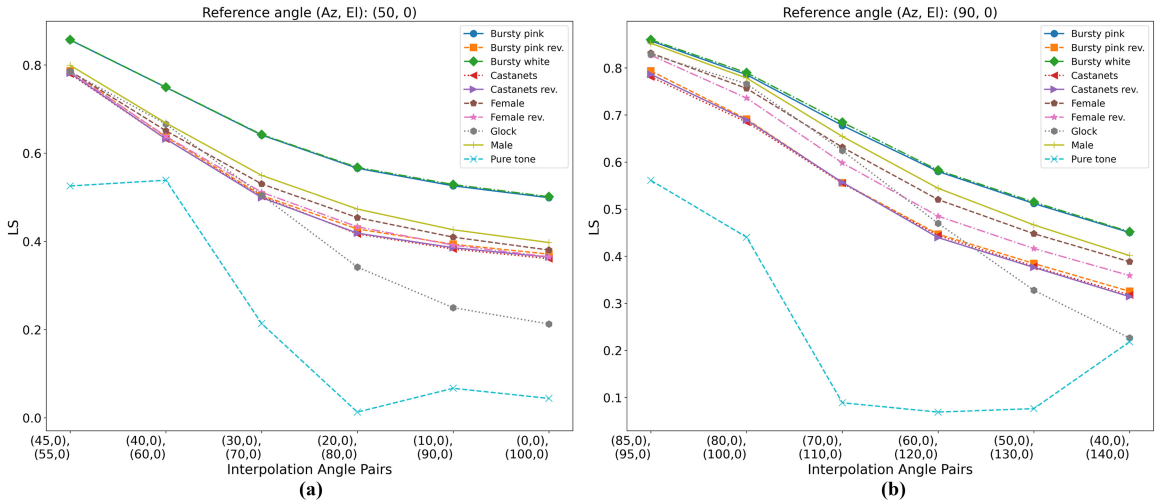


Fig. 4. LS between single-source interpolated test signals in the angle interpolations dataset and their corresponding references. (a) Reference is localized at azimuth = 50°, elevation = 0°. (b) Reference is localized at azimuth = 90°, elevation = 0°.

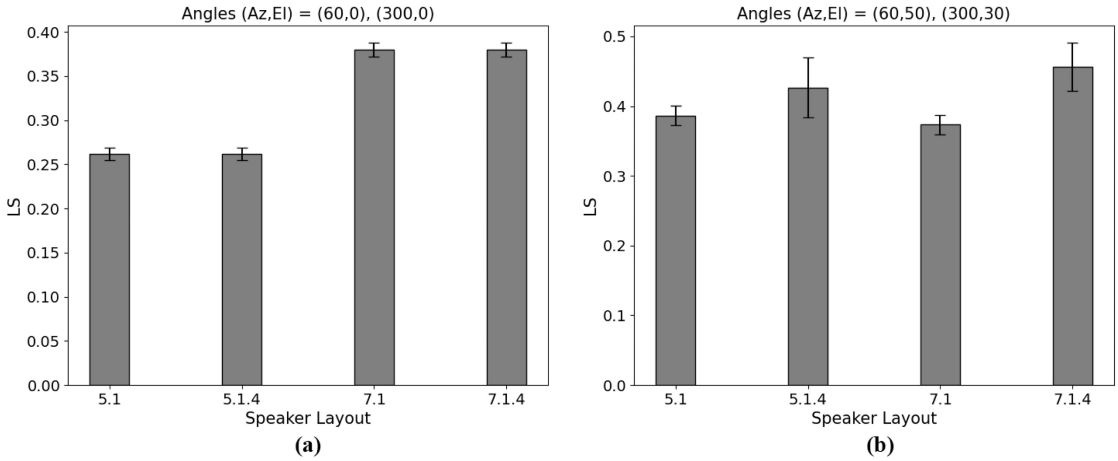


Fig. 5. The average LS between multipoint source samples of the surround layouts dataset, rendered for different loudspeaker layouts, and corresponding reference signals without any layouts. (a) Point source angles without elevation and (b) point source angles with elevation.

Table 6. Statistical comparison between different surround loudspeaker layouts using Mann–Whitney and Kruskal–Wallis tests.

Effect of Increasing the Layout Size		
Comparison	Test	p Value
5.1 vs. 7.1	Mann–Whitney U	0.0002
5.1.4 vs. 7.1.4	Mann–Whitney U	0.0002
All layouts	Kruskal–Wallis	0.0000
Effect of Adding More Height Channels		
Comparison	Test	p Value
5.1 vs. 5.1.4	Mann–Whitney U	0.0499
7.1 vs. 7.1.4	Mann–Whitney U	0.0003
All layouts	Kruskal–Wallis	0.0008

nificant differences due to layout configuration. These results demonstrate that BINAQUAL can distinguish between simulated binaural representations of different surround loudspeaker layouts, regardless of whether elevation is present.

5.4 Codec Compression

LS was computed between all single-point and multiple-point samples of the codec compression dataset and their corresponding references (noncompressed signals). Fig. 6 compares the aggregated LS scores with the results of a subjective listening test performed in [16] with the same set of recordings.

Figs. 6(a) and 6(b) show the aggregated mean values of subjective MUSHRA results across various encodings for single-point and multiple-point source scenarios, respectively. Figs. 6(c) and 6(d) present the BINAQUAL LS results for different encodings in single-source and multiple-source scenarios, respectively. The 95% confidence intervals are also shown. Similar trends are seen between aggregated subjective results (MUSHRA) and aggregated BINAQUAL LS scores. This holds for both single-point and multipoint source scenarios across different audio contents.

Figs. 6(e) and 6(f) depict the relationship between BINAQUAL LS scores and subjective MUSHRA results across different encodings, while Figs. 6(g) and 6(h) illustrate similar comparisons for different audio contents for both single-point and multiple-point source scenarios. It is observed that the data points have been clustered by encoding, while at the same time, there is a lack of clustering by content.

There is a strong correlation between subjective MUSHRA results and LS scores. For single-point source results, the Pearson and Spearman correlations are 0.85 and 0.84, respectively. For multiple-point source results, these correlations increase to 0.87 and 0.92, respectively, suggesting a slightly stronger monotonic relationship in the multiple-point scenario.

Additionally, the correlation between AMBIQUAL localization accuracy scores and BINAQUAL LS scores was computed. For single-point sources, the Pearson and Spearman correlation coefficients are 0.84 and 0.80, respectively. For multiple-point sources, the correlations increase to 0.93 (Pearson) and 0.96 (Spearman). The results indicate that the BINAQAUL model effectively distinguishes between different compression levels of binaural audio signals while remaining robust to variations in content.

5.5 General Discussion

The results presented in this study demonstrate that the BINAQUAL is effective in assessing spatial LS in binaural audio across various conditions. SEC. 5.1 analyzed the model's sensitivity to source localization and showed that BINAQUAL successfully detects small variations in point-source locations, as evidenced by the decrease in LS scores with increasing angular distance from the reference. However, for angles in the range of $\pm 120^\circ$ to $\pm 180^\circ$, LS scores increased, which could be attributed to front-back confusion where binaural cues are ambiguous

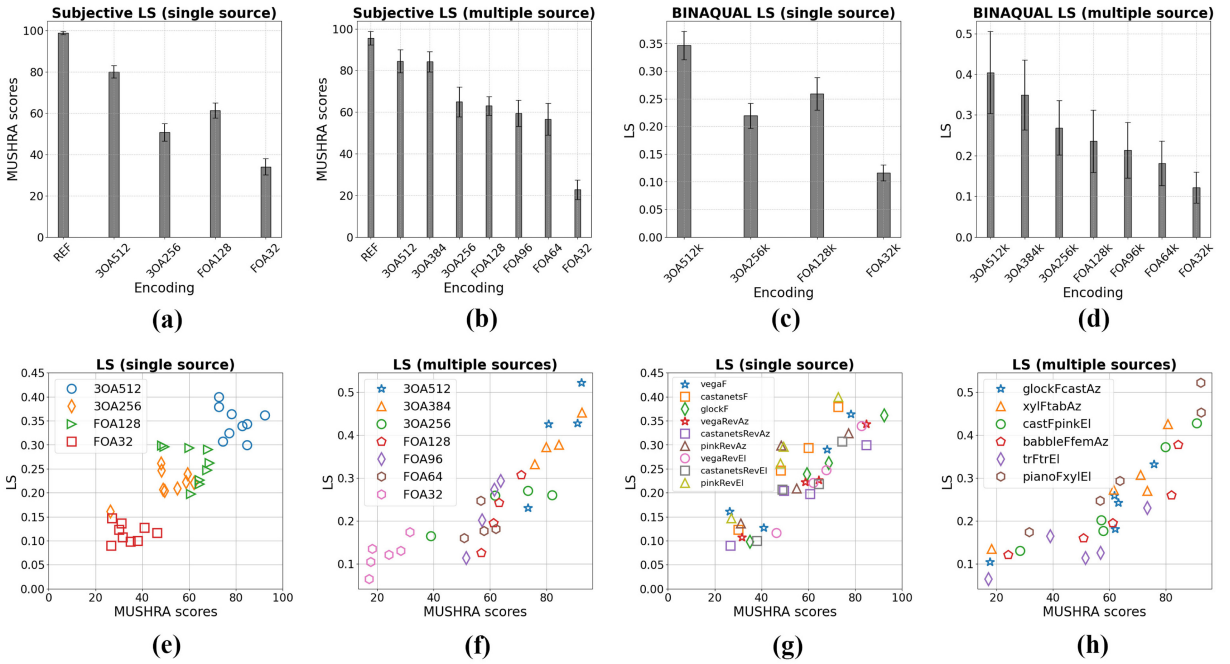


Fig. 6. Subjective LS vs. BINAQUAL results: (a) and (b) show the aggregated subjective MUSHRA test results for different encodings in single-point and multiple-point source scenarios, respectively. (c) and (d) present the aggregated BINAQUAL LS scores for different encodings in single-point and multiple-point source scenarios, respectively. (e) and (f) illustrate the relationship between BINAQUAL LS and MUSHRA scores, grouped by encoding, for single-point and multiple-point source scenarios, respectively. (g) and (h) show the relationship between BINAQUAL LS and MUSHRA scores, grouped by audio content, for single-point and multiple-point source scenarios, respectively.

due to the lack of sufficient spectral differentiation. This observation aligns with psychoacoustic phenomena in binaural hearing [21].

Additionally, it was found that the model is less effective in localizing pure tones due to their lack of rich interaural cues, a limitation that is consistent with human difficulties in localizing narrow-band audio sources [51]. The results also indicate that BINAQUAL maintains relatively flat LS scores at extreme elevations ($\pm 90^\circ$). This behavior is expected, as azimuth resolution effectively collapses at $\pm 90^\circ$ elevation. Also, elevation perception relies more heavily on spectral cues, which are typically more variable and harder to model robustly, further contributing to the observed uniformity in LS scores.

SEC. 5.2 analyzed the model's ability to differentiate between angle interpolations of single-point sources and found that BINAQUAL tracks the decrease in LS as the azimuth angles deviate from the reference position. A similar trend was seen across all contents except pure tone, which again confirms the difficulty of localizing narrow-band signals due to the lack of interaural cues. Similarly, glockenspiel, which has high-frequency spectral content, demonstrated lower LS for more extreme angles, suggesting that a lack of strong ITD cues can reduce localization precision.

SEC. 5.3 showed that BINAQUAL could differentiate effectively between 5.1, 5.1.4, 7.1, and 7.1.4-rendered audio signals, showing that surround loudspeaker layouts with more channels (and especially additional height channels) increase LS. The capability to discriminate among different loudspeaker layouts is particularly promising for applications involving complex multichannel setups in VR, AR, and cinematic audio.

SEC. 5.4 explored the model's ability to distinguish between different compression levels of binaural audio content. BINAQUAL not only differentiated among binaural renders of different compression levels but also correlated strongly with subjective MUSHRA tests. The model achieved higher correlation coefficients in multisource scenarios, indicating that BINAQUAL generalizes well to complex scenes containing multiple sound sources.

Finally, regarding its robustness to variations in content and the number of sources, the model demonstrated stable performance across both single and multipoint source recordings, as well as different audio types (e.g., speech, music, impulsive sounds). This versatility indicates that BINAQUAL is applicable to a wide range of practical scenarios, from simple point-source recordings to more complex, multisource mixes.

While BINAQUAL effectively generalizes across content, source configurations, and degradations, some limitations remain. Its performance declines in conditions where binaural cues are inherently ambiguous or diminished, such as at extreme elevations, with pure tone stimuli, or in cases of front-back confusion. These are also well-known challenges for human listeners, which reinforces the perceptual validity of the metric. It should be noted that the goal is not to outperform human hearing but rather to develop a metric that closely mimics how humans perceive spatial audio quality. Future work could enhance BINAQUAL by

integrating additional spectral cues or simulating listener head movements, strategies that humans naturally use to resolve spatial ambiguities. Such enhancements aim to further align the metric with perceptual realism and improve its applicability to dynamic or immersive spatial audio scenarios.

All five RQs outlined in SEC. 1 were addressed, and the results demonstrated the capabilities of the BINAQUAL model. While most existing spatial localization quality metrics focus primarily on point-source localization accuracy, this study took a broader approach to spatial audio evaluation. By examining the effects of angle interpolations, loudspeaker layouts, codec compression, and content variations, this work provides deeper insights into how spatial fidelity is preserved across diverse scenarios. These aspects are crucial for advancing localization quality assessment models and ensuring objective metrics align more closely with perceptual realism. Consequently, this study not only introduces a novel localization quality assessment metric but also sets a foundation for a more holistic approach to spatial localization quality evaluation beyond traditional accuracy-based methods.

6 CONCLUSION

This study presented a full-reference objective LS metric for binaural audio. Through a comprehensive set of validation experiments, the model's capabilities in differentiating between variations in source positions, angle interpolations, surround loudspeaker layouts, and audio degradations were demonstrated.

The model successfully differentiates subtle spatial variations, as evidenced by its sensitivity to small angular deviations. However, its performance is affected by well-known psychoacoustic limitations, such as front-back confusion and the reduced localization accuracy of pure tones. These findings align with established auditory localization principles, reinforcing the model's validity. Also, BINAQUAL effectively distinguishes between different surround loudspeaker layouts, highlighting its potential for evaluating complex binaural audio formats.

This analysis of compressed binaural audio demonstrates BINAQUAL's strong correlation with perceptual MUSHRA scores, particularly in multisource scenarios. This indicates that the metric can be a reliable proxy for subjective evaluations, offering a scalable and efficient alternative to listener-based assessments. Furthermore, BINAQUAL maintains robust performance across different content types and varying numbers of sound sources, indicating its applicability across diverse spatial audio scenarios.

This work expands spatial audio quality evaluation beyond point-source localization accuracy by considering factors like angle interpolations, loudspeaker layouts, compression artifacts, and content variations, offering a more comprehensive assessment of spatial fidelity in binaural audio. Future work could further refine BINAQUAL's performance to ensure even greater alignment with human spatial perception.

7 ACKNOWLEDGMENT

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under grant numbers 12/RC/2289_P2 and 13/RC/2077_P2 and via a gift to support this research from Google. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

8 REFERENCES

[1] T. Pangarkar, "Smart Wearables Statistics 2025 By Devices, Technology, Usage," https://scoop.market.us/smart-wearables-statistics/ (2025 Jan.).

[2] Google Open Source Blog, "Introducing Eclipsa Audio: Immersive Audio for Everyone," https://opensource.googleblog.com/2025/01/introducing-eclipsa-audio-immersive-audio-for-everyone.html (2025 Jan.).

[3] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," *Recommendation ITU-R BS.1534-1* (2015 Oct.).

[4] J. C. Makous and J. C. Middlebrooks, "Two Dimensional Sound Localization by Human Listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2188–2200 (1990 May). https://doi.org/10.1121/1.399186.

[5] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246 (1958 Apr.). https://doi.org/10.1121/1.1909553.

[6] S. Bech and N. Zacharov, *Perceptual Audio Evaluation–Theory, Method and Application* (Wiley, Chichester, UK, 2007). https://doi.org/10.1002/9780470869253.

[7] A. Ihlefeld and B. G. Shinn-Cunningham, "Effect of Source Spectrum on Sound Localization in an Everyday Reverberant Room," *J. Acoust. Soc. Am.*, vol. 130, no. 1, pp. 324–333 (2011 Jul.). https://doi.org/10.1121/1.3596476.

[8] I. Y. Choi, S. B. Chon, B. G. Shinn-Cunningham, and K.-M. Sung, "Prediction of Perceived Quality in Multi-Channel Audio Compression Coding Systems," in Proc. of the *AES 30th Intl. Conf.: Intelligent Audio Environments* (2007 Mar.), paper 6.

[9] J.-H. Fleßner, R. Huber, and S. Ewert, "Assessment and Prediction of Binaural Aspects of Audio Quality," *J. Audio Eng. Soc.*, vol. 65, no. 11, pp. 929–942 (2017 Nov.). https://doi.org/10.17743/jaes.2017.0037.

[10] J.-H. Fleßner, T. Biberger, and S. D. Ewert, "Subjective and Objective Assessment of Monaural and Binaural Aspects of Audio Quality," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 7, pp. 1112–1125 (2019 Jul.). https://doi.org/10.1109/TASLP.2019.2904850.

[11] K. N. Watcharasupat and A. Lerch, "Quantifying Spatial Audio Quality Impairment," in Proc. of the *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 746–750 (Seoul, South Korea) (2024 Apr.). https://doi.org/10.1109/ICASSP48485.2024.10447947.

[12] J.-H. Fleßner, S. D. Ewert, B. Kollmeier, and R. Huber, "Quality Assessment of Multi-Channel Audio Processing Schemes Based on a Binaural Auditory Model," in Proc. of the *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1340–1344 (Florence, Italy) (2014 May). https://doi.org/10.1109/ICASSP.2014.6853815.

[13] P. Manocha, A. Kumar, B. Xu, et al., "DPLM: A Deep Perceptual Spatial-Audio Localization Metric," in Proc. of the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 6–10 (New Paltz, NY) (2021 Oct.). https://doi.org/10.1109/WASPAA52581.2021.9632781.

[14] P. Manocha, A. Kumar, B. Xu, et al., "SAQAM: Spatial Audio Quality Assessment Metric," in Proc. of the *Interspeech*, pp. 649–653 (Incheon, South Korea) (2022 Sep.). https://doi.org/10.21437/Interspeech.2022-406.

[15] P. Manocha, I. D. Gebru, A. Kumar, D. Markovic, and A. Richard, "Spatialization Quality Metric for Binaural Speech," in Proc. of the *Interspeech*, pp. 5426–5430 (Dublin, Ireland) (2023 Aug.). https://doi.org/10.21437/Interspeech.2023-832.

[16] M. Narbutt, J. Skoglund, A. Allen, et al., "AMBIQUAL: Towards a Quality Metric for Headphone Rendered Compressed Ambisonic Spatial Audio," *Appl. Sci.*, vol. 10, no. 9, paper 3188 (2020 Jan.). https://doi.org/10.3390/app10093188.

[17] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, no. 3–4, pp. 171–218 (1992 Jan.). https://doi.org/10.1016/0003-682X(92)90046-U.

[18] D. R. Moore, P. A. Fuchs, A. Rees, A. Palmer, and C. J. Plack, *The Oxford Handbook of Auditory Science: The Auditory Brain*, vol. 2 (Oxford University Press, New York, NY, 2010).

[19] E. A. Macpherson and J. C. Middlebrooks, "Listener Weighting of Cues for Lateral Angle: The Duplex Theory of Sound Localization Revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2219–2236 (2002 May). https://doi.org/10.1121/1.1471898.

[20] M. Jc, "Sound Localization by Human Listeners," *Ann. Rev. Psychol.*, vol. 42, pp. 135–159 (1991 Feb.). https://doi.org/10.1146/annurev.ps.42.020191.001031.

[21] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997). https://doi.org/10.7551/mitpress/6391.001.0001.

[22] D. R. Begault, "3-D Sound for Virtual Reality and Multimedia," Tech. Rep. (2000 Apr.).

[23] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, paper 2029 (2018 Nov.). https://doi.org/10.3390/app8112029.

[24] M. S. Dobreva, W. E. O'Neill, and G. D. Paige, "Influence of Aging on Human Sound Localization," *J. Neurophysiol.*, vol. 105, no. 5, pp. 2471–2486 (2011 May). https://doi.org/10.1152/jn.00951.2010.

[25] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal, "Relearning Sound Localization With New Ears," *Nature Neurosci.*, vol. 1, no. 5, pp. 417–421 (1998 Sep.). https://doi.org/10.1038/1633.

[26] D. S. Brungart and W. M. Rabinowitz, "Auditory Localization of Nearby Sources. Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479 (1999 Sep.). https://doi.org/10.1121/1.427180.

[27] F. L. Wightman and D. J. Kistler, "Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853 (1999 May). https://doi.org/10.1121/1.426899.

[28] W. M. Hartmann, "Localization of Sound in Rooms," *J. Acoust. Soc. Am.*, vol. 74, no. 5, pp. 1380–1391 (1983 Nov.). https://doi.org/10.1121/1.390163.

[29] S. Kerber and B. U. Seeber, "Sound Localization in Noise by Normal-Hearing Listeners and Cochlear Implant Users," *Ear Hear.*, vol. 33, no. 4, pp. 445–457 (2012 Jul./Aug.). https://doi.org/10.1097/AUD.0b013e318257607b.

[30] R. Ege, A. J. V. Opstal, and M. M. Van Wanrooij, "Accuracy-Precision Trade-off in Human Sound Localisation," *Sci. Rep.*, vol. 8, no. 1, paper 16399 (2018 Nov.). https://doi.org/10.1038/s41598-018-34512-6.

[31] S. Shigeno and T. Oyama, "Localization of Speech and Non-Speech Sounds," *Japan. Psychol. Res.*, vol. 25, no. 2, pp. 112–117 (1983 Sep.).

[32] W. Chau and R. Duda, "Combined Monaural and Binaural Localization of Sound Sources," in Proc. of the *29th Asilomar Conf. on Signals, Systems and Computers*, vol. 2, pp. 1281–1285 (Pacific Grove, CA) (1995 Oct.). https://doi.org/10.1109/ACSSC.1995.540905.

[33] E. Borg, C. Bergkvist, and D. Bagger-Sjöbäck, "Effect on Directional Hearing in Hunters Using Amplifying (Level Dependent) Hearing Protectors," *Otol. Neurotol.*, vol. 29, no. 5, pp. 579–585 (2008 Aug.). https://doi.org/10.1097/MAO.0b013e318172cf70.

[34] J. C. Middlebrooks, "Narrow-Band Sound Localization Related to External Ear Acoustics," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2607–2624 (1992 Nov.). https://doi.org/10.1121/1.404400.

[35] R. A. Butler and R. A. Humanski, "Localization of Sound in the Vertical Plane With and Without High-Frequency Spectral Cues," *Percept. Psychophys.*, vol. 51, no. 2, pp. 182–186 (1992 Feb.). https://doi.org/10.3758/bf03212242.

[36] F. L. Wightman and D. J. Kistler, "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661 (1992 Mar.). https://doi.org/10.1121/1.402445.

[37] W. A. Yost and X. Zhong, "Sound Source Localization Identification Accuracy: Bandwidth Dependencies," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2737–2746 (2014 Nov.). https://doi.org/10.1121/1.4898045.

[38] P. Zakarauskas and M. S. Cynader, "A Computational Theory of Spectral Cue Localization," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1323–1331 (1993). https://doi.org/10.1121/1.408160.

[39] E. H. Langendijk and A. W. Bronkhorst, "Contribution of Spectral Cues to Human Sound Localization," *J.*

Acoust. Soc. Am., vol. 112, no. 4, pp. 1583–1596 (2002 Sep.). https://doi.org/10.1121/1.1501901.

[40] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory Model Based Direction Estimation of Concurrent Speakers From Binaural Signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605 (2011 May/Jun.). https://doi.org/10.1016/j.specom.2011.01.004.

[41] T. May, S. Van De Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 1–13 (2011 Jan.). https://doi.org/10.1109/TASL.2010.2042128.

[42] W. Lindemann, "Extension of a Binaural Cross-Correlation Model by Contralateral Inhibition. I. Simulation of Lateralization for Stationary Signals," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622 (1986 Dec.). https://doi.org/10.1121/1.394325.

[43] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014 Aug.). https://doi.org/10.1121/1.4887447.

[44] H. Ren, C. Ritz, J. Zhao, X. Zheng, and D. Jang, "A Metric for Predicting the Quality of Ambisonic Spatial Audio Reproduced Using Spatially Interpolated or Extrapolated Room Impulse Responses," in Proc. of the *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (Hyderabad, India) (2025 Apr.), paper 7220. https://doi.org/10.1109/ICASSP49660.2025.10889495.

[45] Y. Zheng, J. Yao, X. Deng, et al., "HAPG-SAQAM: Human Auditory Perception Guided Spatial Audio Quality Assessment Metric," in Proc. of the *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (Hyderabad, India) (2025 Apr.), paper 7217. https://doi.org/10.1109/ICASSP49660.2025.10889173.

[46] A. Hines, E. Gillen, D. Kelly, et al., "ViSQOLAudio: An Objective Audio Quality Metric for Low Bitrate Codecs," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL449–EL455 (2015 May). https://doi.org/10.1121/1.4921674.

[47] A. Hines and N. Harte, "Speech Intelligibility Prediction Using a Neurogram Similarity Index Measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320 (2012 Feb.). https://doi.org/10.1016/j.specom.2011.09.004.

[48] D. Barry, D. Shariat Panah, A. Ragano, J. Skoglund, and A. Hines, "Binamix - A Python Library for Generating Binaural Audio Datasets," presented at the *158th Conv. of the Audio Engineering Society* (2025 May). https://doi.org/10.48550/arXiv.2505.01369.

[49] Broadcasting Union European, "Sound Quality Assessment Material," EBU 3253-E (1988 Apr.).

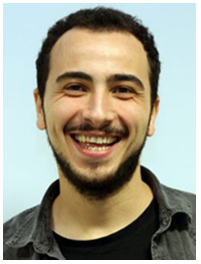
[50] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[51] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Emerald Group, Bingley, UK, 2012), 6th ed.

THE AUTHORS


Davoud Shariat Panah


Dan Barry


Alessandro Ragano


Jan Skoglund


Andrew Hines

Davoud Shariat Panah is a postdoctoral research fellow at the School of Computer Science, University College Dublin, Ireland. He earned a Ph.D. in Computer Science from Technological University Dublin in 2023. His research focuses on audio and speech processing, as well as machine learning applications in healthcare.

•

Dan Barry is a software engineer and product manager specializing in digital audio technology. He holds a Ph.D. in audio signal processing and has held academic positions at Dublin Institute of Technology and University College Dublin. With over 20 years of experience, he has published extensively, holds multiple patents, and cofounded several audio technology companies, including Riffstation, later acquired by Fender. His research covers sound source separation, spatial audio, music information retrieval, audio quality metrics, and audio enhancement algorithms. He has collaborated on technology development with Sony, Google, and other leading audio companies, and continues to drive innovation in both industry and academia.

•

Alessandro Ragano is a postdoctoral research fellow with the School of Computer Science at University College Dublin, where he also completed his Ph.D. in 2022. He earned an M.Sc. in computer engineering from Politecnico Di Milano in 2018. His research interests lie in the application of deep learning techniques to audio signal processing and multimedia quality of experience.

•

Jan Skoglund received a Ph.D. degree from the Chalmers University of Technology, Gothenburg, Sweden, in 1998. He leads a team with Google, San Francisco, CA, developing speech and audio signal processing components for capture, real-time communication, storage, and rendering. These components were deployed in Google software products, such as Meet, and hardware products such as Chromebooks. After receiving his Ph.D. degree, he worked on low-bit-rate speech coding with AT&T Labs-Research, Florham Park, NJ. From 2000 to 2011, he was with Global IP Solutions (GIPS), San Francisco, CA, working on speech and audio processing, such as compression, enhancement, and echo cancellation, tailored for packet-switched networks. GIPS's audio and video technology was found in many deployments, such as IBM, Google, Yahoo, WebEx, Skype, and Samsung, and was open-sourced as WebRTC after a 2011 acquisition by Google. Since then, he has been a part of Chrome with Google.

•

Andrew Hines (AES Member and IEEE Senior Member) is a faculty member with the School of Computer Science, University College Dublin, Ireland. He leads the QxLab Research Group. His research interests include machine perception, multimedia quality of experience, and machine learning for applications in speech, audio, and video signal processing. He developed the ViSQOL speech quality metric and the Go Listen subjective testing platform. He is a member of the Royal Irish Academy Engineering and Computer Sciences Committee.