



Title	AMBIQUAL - a full reference objective quality metric for ambisonic spatial audio
Authors(s)	Narbutt, Mirosław, Allen, Andrew, Skoglund, Jan, Chinen, Michael, Hines, Andrew
Publication date	2018-09-13
Publication information	Narbutt, Mirosław, Andrew Allen, Jan Skoglund, Michael Chinen, and Andrew Hines. "AMBIQUAL - a Full Reference Objective Quality Metric for Ambisonic Spatial Audio." IEEE, September 13, 2018. https://doi.org/10.1109/qomex.2018.8463408 .
Conference details	The 2018 Tenth International Conference on Quality of Multimedia Experience (QoMex), Sardinia, Italy, 29 May- 1 June 2018
Publisher	IEEE
Item record/more information	http://hdl.handle.net/10197/10458
Publisher's statement	© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Publisher's version (DOI)	10.1109/qomex.2018.8463408

Downloaded 2026-05-02 00:29:07

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



© Some rights reserved. For more information

AMBIQUAL – a full reference objective quality metric for ambisonic spatial audio

Mirosław Narbutt
School of Computing
Dublin Institute of Technology
Dublin, Ireland
mirosław.narbutt@dit.ie

Andrew Allen, Jan Skoglund, Michael Chinen
Google, Inc.
San Francisco, Ca., U.S.A.
bitllama, jks, mchinen@google.com

Andrew Hines
School of Computer Science
University College Dublin
Dublin, Ireland
andrew.hines@ucd.ie

Abstract—Streaming spatial audio over networks requires efficient encoding techniques that compress the raw audio content without compromising quality of experience. Streaming service providers such as YouTube need a perceptually relevant objective audio quality metric to monitor users’ perceived quality and spatial localization accuracy. In this paper we introduce a full reference objective spatial audio quality metric, AMBIQUAL, which assesses both Listening Quality and Localization Accuracy. In our solution both metrics are derived directly from the B-format Ambisonic audio. The metric extends and adapts the algorithm used in ViSQOLAudio, a full reference objective metric designed for assessing speech and audio quality. In particular, Listening Quality is derived from the omnidirectional channel and Localization Accuracy is derived from a weighted sum of similarity from B-format directional channels. This paper evaluates whether the proposed AMBIQUAL objective spatial audio quality metric can predict two factors: Listening Quality and Localization Accuracy by comparing its predictions with results from MUSHRA subjective listening tests. In particular, we evaluated the Listening Quality and Localization Accuracy of First and Third-Order Ambisonic audio compressed with the OPUS 1.2 codec at various bitrates (i.e. 32, 128 and 256, 512kbps respectively). The sample set for the tests comprised both recorded and synthetic audio clips with a wide range of time-frequency characteristics. To evaluate Localization Accuracy of compressed audio a number of fixed and dynamic (moving vertically and horizontally) source positions were selected for the test samples. Results showed a strong correlation (PCC=0.919; Spearman=0.882 regarding Listening Quality and PCC=0.854; Spearman=0.842 regarding Localization Accuracy) between objective quality scores derived from the B-format Ambisonic audio using AMBIQUAL and subjective scores obtained during listening MUSHRA tests. AMBIQUAL displays very promising quality assessment predictions for spatial audio. Future work will optimise the algorithm to generalise and validate it for any Higher Order Ambisonic formats.

Keywords—virtual reality, spatial audio, ambisonics, audio coding, audio compression, opus codec, MUSHRA

I. INTRODUCTION

Progress and adoption of emerging technology in the fields of virtual and augmented reality has reinvigorated interest in spatial audio. It is now supported in implementations of headsets like Oculus and by streaming services including YouTube [1]. Audio codecs support efficient streaming and storage of spatial audio are active in the standards community, e.g. with Opus [2] and MPEG-H [3]. Ambisonics is a form of spatial audio that can be stored in B-format [4].

Ambisonics offers a possibility to represent three-dimensional sound in the form of a soundscape, independent of a specific loudspeaker set-up. First Order Ambisonics (FOA) audio is encoded into 4 channels (omnidirectional gain, left/right, up/down, front/back). As found by [5] and validated in [6], with 16 channels, Third-Order Ambisonics (3OA) significantly improves the Quality of Experience (QoE) at the expense of a large amount of data.

Efficient delivery of spatial audio for streaming services with limited bandwidth using Higher-Order Ambisonics has driven development of novel compression techniques, e.g. [7]. Delivering streaming spatial audio requires compression due to bandwidth limitations. In order to measure the perceptual QoE for spatial audio using compressed ambisonics, quality assessment methodologies are required. Subjective tests for spatial audio require even more time and effort than regular speech or audio testing methods (e.g. P.800 [8] or MUSHRA [9]) as both the sound quality and the localisation accuracy of the signal need to be assessed [6].

No objective models for machine-based prediction of spatial audio quality currently exist. Unlike existing metrics for speech or regular audio quality assessment, spatial audio needs to provide an assessment of QoE that takes into account not only the effects of audio fidelity degradations but also whether compression has altered the perceived localization of sound source origins.

This paper presents initial work on adapting an objective audio quality metric to assess the Listening Quality and Localization Accuracy of compressed B-format Ambisonic signals.

This remainder of this paper is organised as follows. It introduces the methods for evaluation of Listening Quality and Localization Accuracy. It gives a brief description of ambisonics and the B-format used for spatial audio. The objective audio quality metric ViSQOLAudio and associated NSIM similarity measure that are extended in this work are introduced. The proposed AMBIQUAL metric is presented and the methods for computing predictions of Listening Quality and Localization Accuracy are explained in detail. The methodology for developing the metric is described followed by validation of the results against a subjective evaluation experiment. Finally, the results are analysed and on-going work is discussed.

II. BACKGROUND

A. Listening Quality and Localization Accuracy

There are a multitude of recommended methodologies for assessing speech and audio quality using subjective listening tests (e.g. ITU Rec's ITU-T P.800 for speech [8], ITU-R Rec. BS.1534-3 [9] and BS.1116-3 [10], and the recently published P.1310 for Spatial audio meetings quality [11]). Objective metrics exist for speech (POLQA [12]) and audio quality (PEAQ [13]) but no objective metrics are agreed upon for spatial audio quality evaluation (although work was began on extending PEAQ [14]). The subjective experiments used in this paper to validate the proposed model were based on binaurally rendered audio presented over headphone so that the compressed ambisonic audio could be evaluated for Listening Quality and Localization Accuracy using the ITU-R Rec. BS.1116 MUSHRA to evaluate Listening Quality and Localization Accuracy compared to a hidden reference. These are described in detail in [6].

Ambisonics can simulate the placement of auditory cues in a virtual 3D space to allow a person's ability to determine the virtual origin of a detected sound. While this has been shown to work, especially using Higher-Order Ambisonics, this paper investigates and presents early work to develop an objective metric that can predict QoE by estimating localization accuracy as well as sound quality.

Extensive study of the mechanisms of auditory localization have been undertaken in the over the last century. Interaural time difference (ITD) and and level/intensity differences (ILD) and the work by Rayleigh [15] on Duplex theory showed that ITD is inferred from phase delays at low frequencies highlighting the relationship between localization direction and phase. Duplex theory asserts that frequencies above 1.5kHz relies on ILD and it has been shown that this is due to the fact that the hair cells in the human cochlea loose the ability to phase lock to higher firing rates. However, more recent research has shown that the human auditory system can use envelope as well as fine structure cues to capture ITD and this has been used in localization models for binaural signals [16].

B. Ambisonics

Ambisonics is a full sphere audio surround technique which is based upon the decomposition of a 3D sound field into a number of spherical harmonics signals. In contrast to channel-based methods with fixed speaker's layouts (e.g. stereo, surround 5.1, surround 7.1) ambisonics contain a speaker-independent representation of a 3D sound field known as B-format, which can be decoded to any speaker layout. This feature is especially useful in Augmented Reality (AR) and Virtual Reality (VR) as it offers good audio signal manipulation possibilities (e.g. rendering audio in real-time according to head movements).

The complete spatial audio information can be encoded into an ambisonics stream containing a number of spherical harmonics signals and scaled to any desired spatial order (an ambisonics stream is said to be of order n when it contains all the signal of orders 0 to n). For example, First-Order Ambisonics (FOA) audio is encoded into 4 spherical harmonics signals: an omnidirectional gain W of order 0 and 3 directional

components of order 1: X (forward/backwards), Y (left/right), and Z (up/down). An ambisonics signal of order 3 contains 16 channels: 1 of order 0, 3 of order 1, 5 of order 2 and 7 of order 3. The number of channels increases with Higher-Order Ambisonics (HOA). Also, the corresponding directional spherical harmonics represent more complex polar patterns allowing more accurate source localization as ambisonics order increases.

Figure 1 illustrates *spherical harmonics* up to third order, sorted by increasing Ambisonic Channel Number (ACN) and aligned for symmetry. The relevant spherical harmonics functions, which provide the direction-dependent amplitudes of each of the ambisonics signals are defined in in Table I.

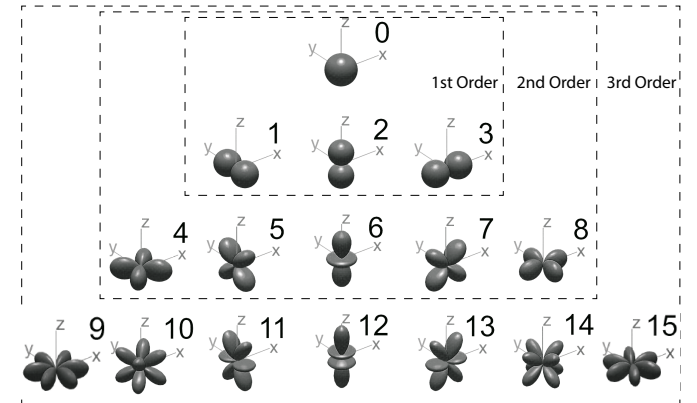


Fig. 1. Visual representation of Ambisonic B-format spherical harmonics signals up to third order. It can be seen that ACN channels 2, 6 and 12 contain only vertical components.

TABLE I. AMBISONICS (FIRST TO THIRD ORDER) EXPRESSING AMPLITUDES AS A FUNCTION OF AZIMUTH (a) AND ELEVATION (e)

ACN (Order)	Formula	ACN (Order)	Formula
0 (0)	1	8 (2)	$\frac{\sqrt{3}}{2} \cos(2a) \cos^2(e)$
1 (1)	$\sin(a) \cos(e)$	9 (3)	$\sqrt{\frac{5}{8}} \sin(3a) \cos^3(e)$
2 (1)	$\sin(e)$	10 (3)	$\frac{\sqrt{15}}{2} \sin(2a) \sin(e) \cos^2(e)$
3 (1)	$\cos(a) \cos(e)$	11 (3)	$\sqrt{\frac{3}{8}} \sin(a) \cos(e) (5 \sin^2(e) - 1)$
4 (2)	$\frac{\sqrt{3}}{2} \sin(2a) \cos^2(e)$	12 (3)	$\frac{1}{2} \sin(e) (5 \sin^2(e) - 3)$
5 (2)	$\frac{\sqrt{3}}{2} \sin(a) \sin(2e)$	13 (3)	$\sqrt{\frac{3}{8}} \cos(a) \cos(e) (5 \sin^2(e) - 1)$
6 (2)	$\frac{1}{2} (3 \sin^2(e) - 1)$	14 (3)	$\frac{\sqrt{15}}{2} \cos(2a) \sin(e) \cos^2(e)$
7 (2)	$\frac{\sqrt{3}}{2} \cos(a) \sin(2e)$	15 (3)	$\sqrt{\frac{3}{8}} \cos(3a) \cos^3(e)$

Moving to HOA improves QoE through improved localization accuracy and in some circumstances the Listening Quality [6]. The downside to HOA is the large amount of processing power required to transform ambisonics multichannel streams into a rendered soundscape. Also, streaming ambisonic data over the networks requires efficient encoding techniques that compress the raw audio content in real time and without significantly compromising QoE. Streaming service providers such as YouTube need a perceptually relevant objective audio quality metric to monitor users' perceived quality and spatial localization accuracy.

C. ViSQOL and NSIM

ViSQOL Speech [17] and ViSQOLAudio [18] are full reference objective metrics for measuring speech quality and audio quality respectively. They are both based on using

NSIM [19], a similarity measure that compares the similarity of signals by aligning and evaluating the similarity across time and frequency bands using a spectrogram-based comparison. ViSQOLAudio calculates magnitudes of the reference and test spectrograms using a 32-band Gammatone filter bank (with lowest frequency 50 Hz and highest frequency 20,084 Hz) to compare their similarity. ViSQOL Speech and ViSQOLAudio also carry out preprocessing of the test signal (i.e. time alignment and level adjustments) to match timing and power characteristics of the reference signal. After pre-preprocessing, the signals are compared with the NSIM similarity metric [19].

III. AMBIQUAL DEVELOPMENT

A. AMBIQUAL Model

We propose a spatial audio quality assessment method, AMBIQUAL, which can be used to assess Listening Quality and Localization Accuracy of spatial audio. The model builds on an adaptation of the ViSQOLAudio algorithm. It predicts perceived quality and spatial localization accuracy by computing signal similarity directly from the B-format Ambisonic audio streams. As with ViSQOLAudio, the model derives a spectro-temporal measure of similarity between a reference and test audio signal. AMBIQUAL derives Listening Quality and Localization Accuracy metrics directly from the B-format Ambisonic audio channels unlike other existing methods that evaluate binaurally rendered signals, e.g. [16]. The aim is to predict a composite QoE for the spatial audio signal that is not focused on a particular listening direction or a given head related transfer function (HRTF) used in rendering the binaural signal.

Spectrograms of the reference and test signals are computed using a short-time Fourier transform (STFT) for each ambisonic channel. In contrast to ViSQOLAudio, AMBIQUAL compares the similarity of the phase angles derived from the reference and test signal spectrograms. Early experiments using intensity level differences yielded inconsistent results. Hence, in accordance with the relationship between localization based on ITD/phase, the filter bank output was used to create a spectrogram of phase angles rather than magnitude levels and these were used for the signal similarity comparisons. Consequently, the STFT with 1536-point Hamming window (50% overlap) is applied to the reference and test signals (using MATLAB built-in functions `spectrogram()` and `angle()`) to create reference and test “phaseograms”. The reference phaseogram is segmented into patches and each reference patch is matched with the most similar test patch using NSIM. Finally, the similarity of each most similar patch pair is calculated across all frequency bands.

The ACN component corresponds to channel index as $k = \text{ACN}$. Making the assumption that the omni-directional channel, $k = 0$, contains a composite of the directional channels, the content of this channel will be representative of the perceptual quality degradations (e.g. due to encoding artefacts but not localization differences). Hence, Listening Quality, LQ , is computed by applying the modified ViSQOLAudio algorithm to the phaseograms of the reference, r and test, t , to channel $k = 0$, i.e.

$$LQ = V(r_0, t_0). \quad (1)$$

LQ is a bounded similarity score between 0 and 1 where 1 is a perfect match.

Localization Accuracy, (LA), is computed as a weighted sum of similarity between reference and test Ambisonic channels. They are grouped into vertical-only and mixed direction channels. For third order, channels $k = 2, 6$ and 12 are vertical only. Generalising for higher orders channels, vertical-only channels, $k_{vertical}$, for order n are,

$$k_{vertical}(n) = n(n + 1). \quad (2)$$

Localization Accuracy (LA) is computed as a weighted sum of similarity between reference, r , and test, t , as follows:

$$LA = \frac{\alpha}{N_{vert}} \sum_{k_{vert}} V(r_k, t_k) + \frac{(1 - \alpha)}{N_{mixed}} \sum_{k_{mixed}} V(r_k, t_k) \quad (3)$$

where α controls a trade-off between the importance of vertical and horizontal Ambisonic components. Vertical channel similarity is emphasised as α increases and is used to control the perceptual localization importance of B-format channels.

B. AMBIQUAL development and optimisation

A test suite of synthetic Third-Order Ambisonic B-format signals was created in order to investigate the prediction trends of the model. A reference audio source and a number of test ambisonic audio sources were rendered on a sphere, each with fixed azimuth and elevation angle.

The reference and test ambisonic audio samples were generated using a pink noise audio signal of 1 second duration and sampled at 48kHz. The reference ambisonic audio sources were rendered to 22 fixed localizations evenly distributed on a quarter of the sphere. The test ambisonic audio signals were rendered at 206 fixed localizations evenly distributed on the whole sphere (i.e. with 30° horizontal and 10° vertical steps). Finally, the Localization Accuracy was calculated for each combination of the reference and test audio sources.

The Localization Accuracy was evaluated for eleven values of alpha weighting factor (i.e. ranging from 0 to 1 with 0.1 increments). The results were visually inspected to identify the alpha weighting giving a monotonically decreasing localization accuracy as the test source moves further from the reference source. Alpha values in the range 0.7 to 0.9 exhibited the anticipated trends. These values were used in the validation experiment described below and 0.7 was found to provide the best fit.

Figure 2 shows graphical representation of LA predictions (colour mapped in grayscale) on a sphere when the reference audio source was localized at 60° azimuth and 60° elevation. The test Localization Accuracy can be seen to decreasing further from the reference source as the grey dots get darker. Figure 3 presents the LA predictions for the same example reference source, plotting LA values as a function of azimuth (i.e. from -180° to 180° at 30° angle steps) and 18 elevation angles (i.e. from -90° to 90° in 10° angle steps).

We repeated this procedure for 22 fixed reference audio source localizations. These tests showed that LA decreases

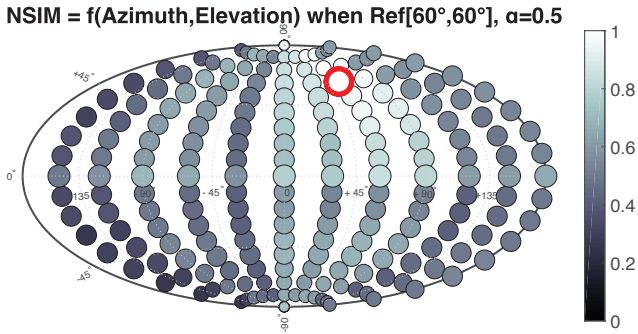


Fig. 2. Graphical representation of Localization Accuracy a sphere when the reference audio source was localized at azimuth=60°, elevation=60°.

monotonically as the test audio source moves away from the reference audio source (both vertically and horizontally) reaching a point of inflection at around +/-90° angle in relation to the audio source localization what corresponds to human ears' localization at +/-90°.

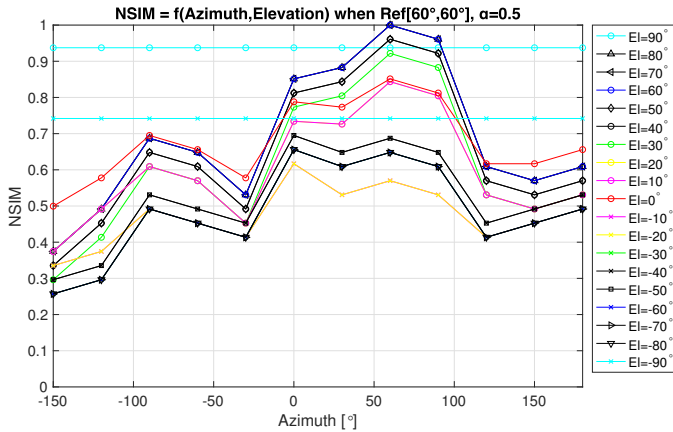


Fig. 3. Localization Accuracy as a function of azimuth and elevation with fixed reference audio source, localized at an offset point of azimuth=60°, elevation=60° (causing asymmetry in results as the source is closer to one ear than the other).

IV. VALIDATION EXPERIMENTS

We evaluated whether the proposed AMBIQUAL objective spatial audio quality metric can predict two factors: Overall Listening Quality and Localization Accuracy, by comparing its predictions with results from MUSHRA subjective listening tests. Full details of the methodology and subjective results are presented in [6]. For clarity, we denote Third-Order Ambisonics as 3OA, rather than HOA, as was used in [6].

Samples were created by converting original stereo samples to mono format and encoding them to FOA and 3OA ambisonic audio for a variety of localizations (i.e. fixed localizations, variable azimuth angle, and variable elevation angle) as shown in Figure 4. The test used 7-15s duration samples (see Table II) for First and Third-Order Ambisonic clips for a range of bitrates (see Table III) and rendered to a binaural format for presentation.

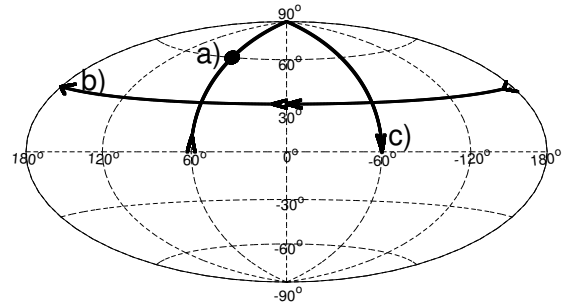


Fig. 4. Localization of sound sources: a) fixed localization (azimuth 60°, elevation 60°), b) dynamic azimuth localization with audio source moving horizontally (i.e. rotating azimuth above the listener's head), c) dynamic elevation localization with audio source moving vertically (i.e. moving up in elevation on the left hand side, then down on the right hand side).

The test conditions were created by encoding ambisonic B-format content using the OPUS 1.2 codec with Channel Mapping Family 2 implementation [7].

TABLE II. AUDIO SAMPLES USED DURING LISTENING TESTS

Label	Music Type	Source
VegaF	Vocals (Suzanne Vega)	CD
CastanetsF	Castanets	EBU
GlockF	Glockenspiel	EBU
vegaReverb	Vocals (Suzanne Vega) w. Reverb	processed CD
CastanetsReverb	Castanets w. Reverb Effect	processed EBU
PinkReverb	Bursty Pink Noise w. Reverb Effect	synthetically generated

TABLE III. LISTENING TEST ENCODING/COMPRESSION SCHEMES

Type	Ambisonics order	Bitrate (kbps)	Bitrate per channel (kbps)
Reference	3	12288	768
3OA 512	3	512	32
3OA 256	3	256	16
FOA 128	1	128	32
FOA 32 (anchor)	1	32	8

AMBIQUAL was used to assess both Listening Quality and Localization Accuracy of the same 6 sample sets, compressed with 4 various bitrates against their uncompressed versions.

Results from the subjective tests and AMBIQUAL's objective predictions are presented for comparison. Figure 5 summarises the aggregated mean values of the Listening Quality and Localization Accuracy for four encoding schemes (i.e. 3OA512, 3OA256, FOA128, and FOA32) and the reference. These aggregated quality scores are shown here as the average MUSHRA score obtained for all 9 audio test samples and the subjective scores' error bars indicate the 95% confidence intervals for four encoding schemes and the reference.

Listening tests showed that the 3OA512 and FOA128 encoding schemes (both 32kbps per channel) perform *good* on the MUSHRA scale in regards to Listening Quality. 3OA512 performs *excellent* in regards to Localization Accuracy what confirms that Third-Order Ambisonics significantly improves the QoE. Lower bitrates per channel (<32kbps) have an adverse impact on QoE. For example, 3OA 256 (16kbps) no longer outperforms FOA128.

As described in the Section III-A, Listening Quality is derived from the B-format ambisonic audio by applying a

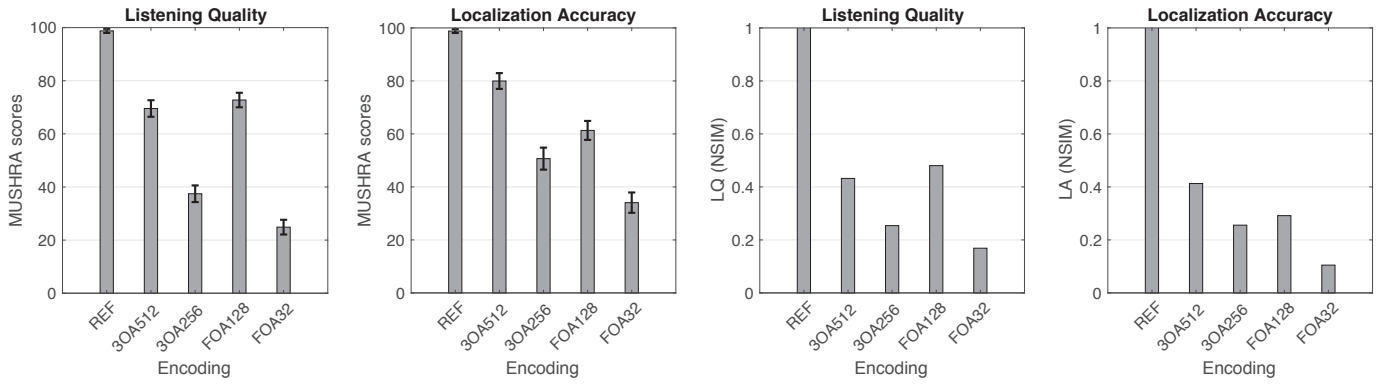


Fig. 5. Listening Quality and Localization Accuracy: Aggregated subjective test results (left two panes) and AMBIQUAL model scores (right two panes) by encoding scheme.

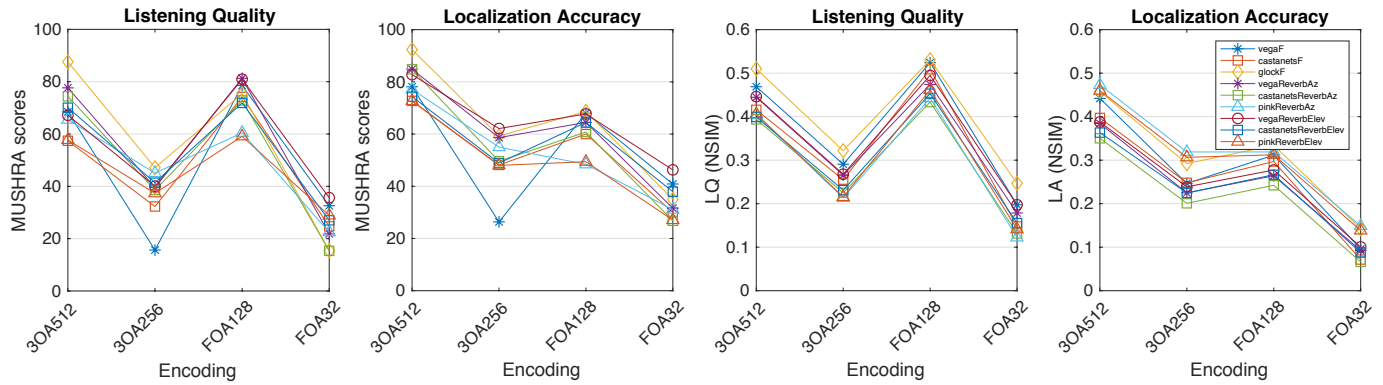


Fig. 6. Listening Quality and Localization Accuracy results from Figure 5 broken down by sample type.

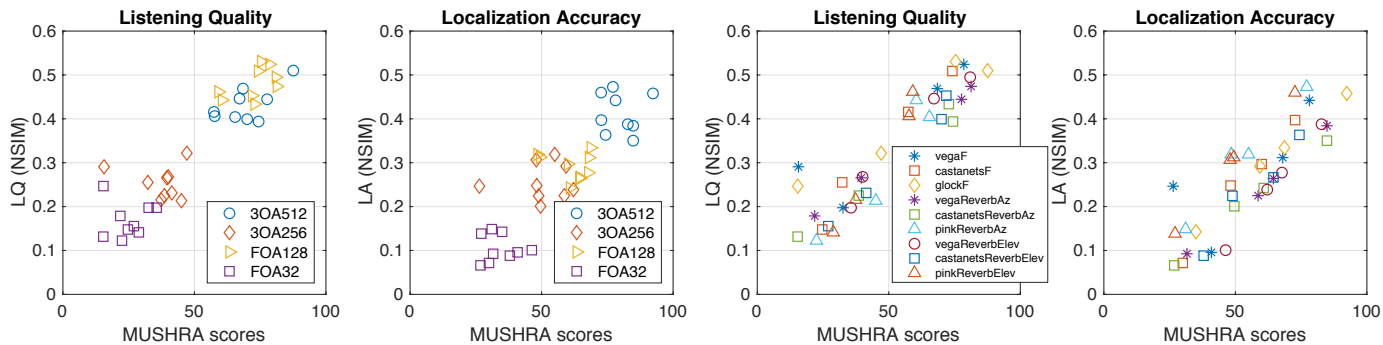


Fig. 7. Scatter of subjective test results versus AMBIQUAL model predictions for Listening Quality and Localization Accuracy. Left panes are by encoding scheme and right side are by sample type.

modified version of the ViSQOLAudio algorithm to the omnidirectional ambisonics channel and Localization Accuracy is computed as a weighted sum of similarity between B-format reference and test directional channels. From the analysis in section III-B a weighting factor $\alpha=0.7$ was chosen to compute the *LA* and *LQ* for the calculations.

A breakdown of the subjective results and the AMBIQUAL model predictions are presented in Figure 6. Both Listening Quality and Localization Accuracy are broken down by encoding scheme for each of the nine audio samples.

The subjective results are compared to the AMBIQUAL

objective model results in Figure 7. Pearson correlation and Spearman rank correlation were computed for the results of all tests excluding comparison of the reference to itself as this would skew the results due to the perfect matches for the objective model yielding higher ranking scores. The results show a strong correlation (PCC=0.919; Spearman=0.882 regarding Listening Quality and PCC=0.854; Spearman=0.842 regarding Localization Accuracy) between objective NSIM scores derived from the B-format ambisonic audio using AMBIQUAL algorithm and subjective MUSHRA scores obtained during listening tests.

V. DISCUSSION AND ON-GOING WORK

AMBIQUAL displays very promising predictive capability for spatial audio QoE regarding both Listening Quality and Localization Accuracy. Results show a strong correlation between objective model predictions and subjective MUSHRA scores obtained during listening tests.

Figure 5 (aggregated quality per encoding scheme) and Figure 6 (aggregated quality per encoding scheme broken down by test samples) show that the trends exhibited in the subjective results are replicated by the objective results. This is true for both Listening Quality and Location Accuracy across all conditions tested.

The clustering of data points by condition (i.e. bit-rate and compression scheme) that is evident for both Listening Quality and Location Accuracy in Figure 7 indicates that the model can predict the difference in quality independently of the test sample content. This observation is reinforced by the lack of clustering by sample (presented in the two scatter plots to the right hand side scatter plots in Figure 7).

As previously stated, the AMBIQUAL model presented in this paper is still at an early stage from a development perspective. The authors are aware of the limitations of the experiments, such as simple testing scenarios (i.e. limited to single point audio sources) and simple encoding scheme used to compress ambisonic signals (i.e. OPUS 1.2, channel mapping 2). A mapping transformation between NSIM similarity (0-1) and MUSHRA (0-100) is also required to replace the current scaling with a perceptually based fitting. Work is ongoing to further test more complex scenarios with a plurality of spatial audio sources to gather more realistic experimental results. Also, more complex compression schemes which share spatial information across the ambisonics channels will be taken into account as they may influence the similarity scores derived from the compressed B-format ambisonics without impacting perceived quality. In addition, new test samples will allow the potential of combining ITD and ILD cues to be investigated with potential further development to the model to deal with these scenarios.

Finally, at a more general level, subjective judgements of QoE regarding how listeners assess Listening Quality and whether it can accurately be judged independently from Location Accuracy is discussed in [6] but is an open question as people may be penalising Listening Quality scores due to Location Accuracy issues. Current research suggests that spatial audio systems can be characterized by spatial (e.g. scene depth, localization accuracy) and non-spatial attributes (e.g. brilliance, distortions). Thus, investigators need to decide whether or not to include non-spatial attributes in quality assessment [20]. Until relevant standards are in place, it is recommended that state-of-the-art methods from multidimensional quality assessment should be applied in order to find an adequate set of dimensions [11].

The simplicity of AMBIQUAL over alternative methods of assessing ambisonic spatial audio such as rendering a range of head positions and carrying out binaural assessment is a very appealing. This computational simplicity combined with the promising early results presented strongly motivates the proposed further research to optimise, generalise and validate it for higher order ambisonic formats.

ACKNOWLEDGEMENT

This publication has emanated from research supported by Google, Inc. and in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

REFERENCES

- [1] J. Brettle and J. Skoglund, "Open-source spatial audio compression for vr content," in *SMPTE 2016 Annual Technical Conference and Exhibition*, Oct 2016, pp. 1–9.
- [2] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," *IETF*, September, 2012. [Online]. Available: RFC6716, <http://www.ietf.org/rfc/rfc6716.txt>
- [3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—the new standard for coding of immersive spatial audio," *IEEE Journal of selected topics in signal processing*, vol. 9, no. 5, pp. 770–779, 2015.
- [4] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.
- [5] S. Bertet, J. Daniel, and E. Parizet, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, pp. 642–657, 2013.
- [6] M. Narbutt, J. Skoglund, A. Allen, and A. Hines, "Streaming VR for immersion: Quality aspects of compressed spatial audio," in *2017 23rd International Conference on Virtual System Multimedia (VSMM)*, Oct 2017.
- [7] J. Skoglund and M. Graczyk, "IETF internet-draft: Ambisonics in an ogg opus container," 2017. [Online]. Available: <http://tools.ietf.org/html/draft-ietf-codec-ambisonics-02>
- [8] ITU, "ITU-R Rec. P.800: Methods for subjective determination of transmission quality," Int. Telecomm. Union, Geneva, 1996.
- [9] ITU, "ITU-R Rec. BS.1534-3: Subjective assessment of sound quality," Int. Telecomm. Union, Geneva, 2015.
- [10] ITU, "ITU-T Rec. BS.1116-3: Methods for the subjective assessment of small impairments in audio systems," Int. Telecomm. Union, Geneva, 2015.
- [11] ITU, "ITU-T Rec. P.1310: Spatial audio meetings quality," Int. Telecomm. Union, Geneva, 2017.
- [12] ITU, "ITU-R Rec. P.863: Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, 2014.
- [13] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ - the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.
- [14] S. Kmpf, J. Liebetrau, S. Schneider, and T. Sporer, "Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, Oct 2010.
- [15] L. Rayleigh, "Xii. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.
- [16] M. Park, P. A. Nelson, and K. Kang, "A model of sound localisation applied to the evaluation of systems for stereophony," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 825–839, 2008.
- [17] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 1, 2015.
- [18] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.
- [19] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306 – 320, 2012.
- [20] N. Zacharov, C. Pike, F. Melchior, and T. Worch, "Next generation audio system assessment using the multiple stimulus ideal profile method," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference*. IEEE, 2016.