EgoMusic: An Egocentric Augmented Reality Glasses Dataset for Music

Alessandro Ragano
University College Dublin
Dublin, Ireland
alessandro.ragano@ucd.ie

Carl Timothy Tolentino
University College Dublin
Dublin, Ireland
carl.tolentino@ucd.ie

Kata Szita
Dublin City University
Dublin, Ireland
kata.szita@dcu.ie

Dan Barry
University College Dublin
Dublin, Ireland
dan.barry1@ucd.ie

Davoud Shariat Panah
University College Dublin
Dublin, Ireland
davoud.shariatpanah@ucd.ie

Niall Murray
Athlone Institute of Technology
Athlone, Ireland
nmurray@research.ait.ie

Andrew Hines
University College Dublin
Dublin, Ireland
andrew.hines@ucd.ie

Abstract

Although audio-augmented reality (AAR) has known applications in music, the use of wearables such as augmented reality (AR) glasses for egocentric audio data capture for music has not been investigated. Current egocentric datasets are mostly focused on speech research, neglecting music's unique demands for tasks such as real-time optimisation or assistive listening. This paper introduces EgoMusic, a multimodal dataset featuring synchronised egocentric audio-visual data captured with AR glasses during live performances, alongside studio-quality audio references. We investigate AR glasses' utility for music and baseline artificial intelligence (AI) approaches for hearing enhancement, positioning EgoMusic as the first dataset that enables research for egocentric music AAR.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; • **Applied computing** → **Sound and music computing**; • **Information systems** → **Multimedia databases**.

Keywords

Audio Augmented Reality, Wearables, Egocentric, Machine Perception

ACM Reference Format:
Alessandro Ragano, Carl Timothy Tolentino, Kata Szita, Dan Barry, Davoud Shariat Panah, Niall Murray, and Andrew Hines. 2025. EgoMusic: An Egocentric Augmented Reality Glasses Dataset for Music. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758262

Figure 1: Musicians during a data capture session for the EgoMusic dataset.

1 Introduction

Audio Augmented Reality (AAR) blends computer-generated sounds with a user's real-world environment, enriching their perception [53]. This technology shows promise in diverse fields, including immersive gaming [7, 30] and interactive education [53].

In music, AAR and visual Augmented Reality (AR) have seen remarkable advances [46]. Research has focused on creating engaging learning environments [27] using visual aids such as fingering guides [13, 22] and 3D models [12, 52]. For live performances, AAR has enabled interfaces for virtual sheet music [23] and for mixing sound loops with spatial audio [31, 46]. Despite progress in AAR for music, using wearables to capture rich, contextualised audio for musical applications is relatively unexplored. Similar to egocentric computer vision [37] using wearable cameras, egocentric audio holds potential. AR glasses are promising platforms for capturing high-fidelity, egocentric audio-visual data due to ergonomics and non-stigmatisation. For instance, AR glasses have been used for speech enhancement [14, 15], audio-visual speaker localization [55], and gaze anticipation [24]. The success of AR glasses and wearables in speech-related applications naturally raises an interesting question of how this technology can be applied or adapted

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758262

for music. Research-focused AR glasses, distinct from consumer devices, are now available. Meta's Aria glasses [17], for example, are used in machine perception research for everyday activities, aiding development of models for assistive technology, robotics, and human-computer interaction [20, 25, 29, 54].

However, AAR's potential as a data acquisition tool for data-driven music models remains largely unexplored, as existing egocentric audio-visual datasets are often not music-focused. This gap allows investigation into using wearable devices such as smart glasses to collect rich data for data-driven models in applications such as real-time music optimisation and remixing.

To address this gap, we introduce EgoMusic: a novel, synchronised egocentric audio-visual dataset from live music performances using Aria glasses. Our primary goals are to investigate AR wearables' potential and limitations for music and explore AI/signal processing baselines for hearing enhancement. The dataset includes aligned, studio-quality reference stems, and this paper details a replicable data collection procedure. EgoMusic is the first egocentric dataset for personalised music enhancement with wearables, serving as a valuable resource for advancing AI-driven, AR-powered music experiences. The egocentric data acquisition methodology presented in this paper promises broader impact across music AAR, including areas such as improved audience accessibility, enhanced music engagement (virtual instrument participation), advanced music education, and novel remote performances.

The key contributions of this paper are: (i) a detailed outline of the challenges of using wearable devices for real-time music optimisation; (ii) a novel egocentric dataset of live music performances including multimodal sensor data such as 7-channel egocentric audio, studio-quality audio references, RGB videos, eye camera videos, magnetometers, and barometers; (iii) a thorough analysis of the Aria glasses' audio fidelity in music recording using objective quality metrics and subjective listening test quality; (iv) an investigation into music remixing using the Aria glasses, demonstrating the potential for personalised music experiences.

2 Motivations and Related Work

Modifying surrounding soundscapes has driven developments from noise-cancelling headphones to hearing aids and transparency modes [48], which mostly offer passive noise control or amplification. In contrast, AAR dynamically manipulates the acoustic environment in real-time, promising personalised music experiences. Traditional music enhancement, typically in post-production, improves pre-recorded audio via numerous signal processing and data-driven techniques such as dynamic range compression, spatialisation, reverberation, equalisation, bandwidth extension [26], denoising [4], remixing [32, 51], and quality restoration [6, 21, 34]. However, adapting such studio-centric techniques to live music wearable AAR is challenging.

Audio enhancement for live scenarios (not pre-recorded music) often has wearable limitations. For instance, Tahmasebi et al. [45] proposed real-time vocal remixing for cochlear implant users, but reliance on virtualised acoustics [18] still limits it to pre-recorded content or specialised hardware, unsuitable for wearables in natural settings. Similarly, real-time classical music source separation (MSS) using multichannel NMF [5] required extensive, non-wearable microphone arrays. Sensor-rich wearables like Meta's Aria glasses [17], capturing egocentric multichannel audio, video, and motion data, offer new pathways. These enable datasets reflecting the wearer's perspective, promoting research into generalizable data-driven AAR solutions beyond lab conditions or pre-recorded stimuli. Such egocentric data is pivotal for advancing real-time music enhancement on wearables in diverse, natural soundscapes.

Despite this potential, wearable AAR's practical realisation, especially for music, faces significant challenges. Parallels with the speech domain highlight common issues, but music presents unique, unaddressed complexities that we outline at the end of this paragraph. Wearable AAR has progressed significantly in speech, mainly targeting source separation (e.g., the cocktail party problem) to enhance communication in noise [15, 19, 40, 43]. Several egocentric speech-focused datasets exist with some limitations. Epic Kitchens [11] lacks multichannel audio and noisy settings. COSINE [44] lacks video/wearable data. EgoCom [28] has limited audio channels and sync issues. Datasets like DiPCo [39], CHiME-5 [2], and CHiME-6 [50] offer rich audio but miss egocentric video and wearable configurations. EasyCom [15] is a significant step, with synchronised multimodal data (audio, video, pose, head-tracking) from AR glasses in realistic conversations, enabling research in speech enhancement, diarization, beamforming, and speech recognition while interacting with the physical world.

Beyond fundamental real-time wearable system challenges (e.g., low latency, data synchronisation, audio overlay - we observe that the lack of speakers in the Aria glasses limits real-time output, but not data acquisition/analysis), music AAR faces distinct challenges from speech that need to be addressed when building an egocentric dataset. These challenges are:

- **Audio Fidelity vs. Speech Intelligibility:** Unlike speech enhancement targeting intelligibility, music AAR demands high *audio fidelity*. This broader, subjective target (clarity, timbral accuracy, spatial realism, dynamic range, perceived quality) is harder to achieve and evaluate.
- **Impact of Distance and Loudness:** Music performance and listening involve wide-ranging distances and sound levels, affecting loudness for listeners. Performers risk microphone clipping from loud sources due to music's wider dynamic range and higher peak levels than speech. Capturing high-quality audio in these conditions with wearable microphones is not trivial.
- **Complexity of Music Source Separation and Remixing:** Unlike speech AAR separating a few speakers, music AAR may need to separate many instruments/vocals in dense mixes. Separated sources must retain high *timbral fidelity* to achieve audio enhancement and personalisation after remixing. MSS algorithm degradations can compromise the augmented experience, so assessing MSS quality on wearable-captured audio is essential. We observe that high-quality source separation and remixing are key for accessibility for hard-of-hearing listeners.
- **Evaluation Metrics:** Standard objective audio metrics may not capture music AAR's perceptual qualities. Effective evaluation needs metrics for spatial accuracy, dynamic range,

timbral preservation, plus subjective assessments (e.g., immersion, enjoyment) [16, 41].
- **Hardware Configuration:** Microphone array configuration (number, type, placement) on wearables directly impacts captured audio quality and spatial algorithm effectiveness (e.g., beamforming, localisation). Understanding this impact on music fidelity is key for hardware/algorithm design.

3 EgoMusic Dataset

As applying egocentric wearable capture to music is an unexplored field, the data collection process and device evaluation must directly address the challenges outlined in Section 2 (e.g., audio fidelity, source complexity). To address the previously outlined lack of suitable data for wearable music AAR systems, we introduce EgoMusic, a novel multimodal dataset[1]. It captures circa one hour of egocentric live music performances over four sessions, simulating realistic listening/performing scenarios at various distances. Our dataset provides synchronised, egocentric multichannel audio-video of live music with high-quality studio reference audio. EgoMusic features data from Aria glasses [17] (worn by performers/audience), reference audio, and 360° video, facilitating research in real-time music source separation, enhancement, spatial audio analysis, and multimodal machine learning for wearable AAR. The next sections detail data collection, equipment, and spatial configuration.

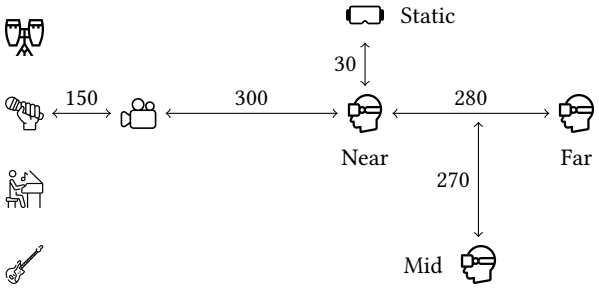


Figure 2: The spatial outline of the recording setup with distances in centimetres.

3.1 Procedure and Participants

Ethical approval was obtained from the UCD Human Research Ethics Committee, and all participants gave informed consent and signed data release agreements. Based on pilot sessions, we recruited one female vocalist and four instrumentalists (acoustic guitar, piano, bass guitar, percussion). To capture a range of musical complexities and listening perspectives, directly addressing the challenges of source separation in dense mixes and the impact of listener position, four distinct recording sessions were conducted:

- **Session 1 (Audience):** 3 songs performed by vocalist + 1 instrument (acoustic guitar or piano).
- **Session 2 (Audience):** 4 songs performed by vocalist + 2 instruments (acoustic guitar/piano + bass guitar).
- **Session 3 (Audience):** 4 songs performed by vocalist + 3 instruments (acoustic guitar/piano + bass guitar + percussion).
- **Session 4 (Musician):** 2 songs performed by vocalist + 1 instrument (acoustic guitar or piano), with glasses worn by the performers.

The total duration of the recorded performances is 58 minutes and 20 seconds. A clap sound at the beginning of each session serves as a synchronisation marker across all independent recording devices. The provided data streams have been manually pre-synchronised using the clap markers.

3.2 Collected Data

Data was captured by recording performances of copyright-free music with multiple synchronised devices (see Table 1).

Wearable Aria Glasses. Four pairs of Aria glasses were used within the same session and distributed between the audience and musicians based on the session details. We used Aria's *Profile 0* (full sensor suite), which include:

- **Audio:** 7-channel @ 48 kHz (5 front-facing mics, 1 near each temple (see Fig 3). Provided tracks are downsampled to 44.1 kHz (matching reference audio).
- **Video:** RGB camera (2880x2880 resolution, with 1 frame rate per second (fps) due to Profile 0 limits, FoV 110°x110°); two monochrome SLAM cameras (640x480, 10 fps, FoV 150°x120°).
- **IMU & Eye Tracking:** IMU @ 1 kHz (left)/800 Hz (right); Eye camera @ 320x240, 10 fps.

For audience sessions (1-3), three volunteers wore glasses at predefined locations (Sec. 3.3); a fourth pair was static on a stand (baseline without head motion).

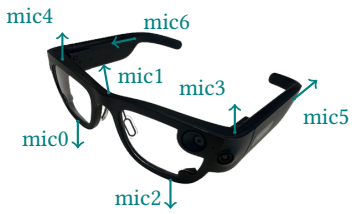


Figure 3: The microphone locations and orientations on the Aria glasses.

Reference Audio. To enable robust evaluation of audio fidelity and provide ground truth for music source separation algorithms, high-quality reference audio was captured for each sound source independently. Condenser microphones were used for close-miking the vocalist, piano, and percussion. The acoustic guitar and bass guitar were recorded via direct input (DI). All reference signals were routed through a Presonus 1812 audio interface and recorded as separate, clean tracks at 44.1 kHz and 24-bit resolution in WAV format. A stereo mix of these reference tracks is also provided for each song.

Spatial Video. To acquire reference video streams, an Insta360 camera, positioned approximately 1.5 meters in front of the musicians (see Figure 2), recorded 360-degree panoramic video footage of the sessions in 4K resolution.

[1] https://doi.org/10.5281/zenodo.16753794

Table 1: Overview of Recording Devices and Data Streams Included in the Dataset

Device / Sensor	Target / Purpose	Key Specifications (Data in Dataset)
Egocentric Wearable System		
Aria Glasses	Egocentric perspective capture	Profile 0 settings
	7-Microphone Array	7-channel spatial audio (44.1 kHz, 24-bit)
	RGB Camera	1 video stream (2880x2880 px, 1 fps, 110° Horiz. FoV)
	SLAM Cameras (x2)	2 B&W video streams (640x480 px, 10 fps, 150° Horiz. FoV)
	IMU (Accelerometer + Gyro)	Motion tracking: 1 kHz (L), 800 Hz (R)
	Eye Tracking Cameras (x2)	Gaze tracking video (320x240 px, 10 fps)
Reference Audio Capture		
Microphones & DI	Clean source recording	Multi-track via Presonus Studio 1810c (.wav, 44.1 kHz)
	Close Mic (Vocals)	Condenser microphone
	Close Mic (Piano)	Condenser microphone
	Close Mic (Percussion)	Condenser microphone
	DI (Acoustic Guitar)	Clean instrument line signal
	DI (Bass Guitar)	Clean instrument line signal
Panoramic Video Capture		
Insta360 Camera	360° performance space video	4K wide-angle video

3.3 Spatial Outline

Recordings took place in a black box theatre with approximate stage dimensions of 7m (L) x 5m (W) x 5m (H). Musicians were positioned adjacent to each other on one side. The physical setup was designed to reflect realistic scenarios and address challenges associated with the impact of distance, loudness, and spatial perception in music AAR (Section 2). Specifically, audience members wearing Aria glasses were situated at varying distances and angles, as depicted in Figure 2. The 'static' Aria head was positioned alongside the audience volunteers.

4 Experiments on the Dataset

To highlight the potential of the glasses to capture high-fidelity live music, objective and subjective audio quality experiments were conducted on sessions 1–3 of EgoMusic. Music source separation was also explored to demonstrate its potential for personalised music streaming. These experiments aim to answer the following research questions: (RQ1) How does the distance of the Aria glasses from the sound source impact the audio quality? (RQ2) How does the number of microphones used in the Aria glasses impact the audio quality? (RQ3) How well does music source separation work on the Aria glasses recordings?

4.1 Audio Quality Objective Tests

VISQOL [8, 42] was used to evaluate the objective audio quality of the recordings. It is used to predict the mean opinion score (MOS) of an audio signal, i.e., the Aria recordings, compared against a reference, i.e., the clean recordings.

To satisfy the input signal requirements, the reference stereo recordings were resampled to 48 kHz, and converted into mono by averaging the left and right channels. Five 10-second audio samples were extracted from five equidistant points from start to finish of each performance, and 0.5-second silence frames were appended at the start and end of each sample. EBU R128 loudness normalisation was employed on all audio samples [47]. In total, 55 audio samples were acquired for each group, categorised according to location (near, mid, far, and static) and number of microphones (one, three, five, and seven).

For multi-microphone evaluation, the delay-and-sum beamforming from SpeechBrain [35, 36] was employed. The microphone combinations used were the following: (one) *mic1*; (three) *mic1* and *mic5–6*; (five) *mic1* and *mic3–mic6*; and (seven) *mic0–mic6*.

Table 2 and Fig. 4 shows the results of the objective tests. One-way analysis of variance (ANOVA) revealed that the MOS was significantly affected by the number of microphones and the location of the glasses. In particular, employing beamforming on the glasses, even for at least three microphones, significantly improved the MOS. However, Tukey's honestly significant difference test (Tukey-HSD) revealed that increasing the number of microphones from three to seven did not significantly improve the MOS.

Table 2: The average MOS and the p-value according to location and number of microphones

Aria	1 mic	3 mics	5 mics	7 mics	p_{mic}
near	2.82	3.19	3.28	3.28	**< 0.01**
mid	2.73	2.98	3.02	2.91	**0.03**
far	2.64	2.94	3.04	3.05	**< 0.01**
static	2.79	3.10	3.12	3.17	**< 0.01**
p_{loc}	0.34	**0.04**	**0.04**	**< 0.01**	

Meanwhile, MOS was significantly affected according to location when beamforming was applied. Based on Tukey-HSD, the Aria-mid recordings had the worst MOS. Also, the MOS was not significantly affected when comparing the recordings from the human-worn glasses and the static glasses.

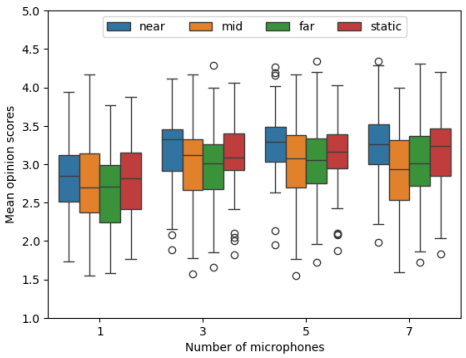
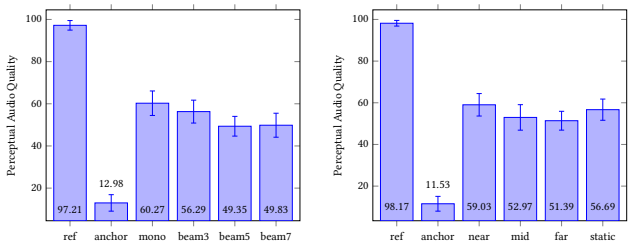


Figure 4: The VISQOL-based MOS on the Aria glasses recordings according to the number of microphones and location.



(a) Microphones **(b) Location**

Figure 5: The MUSHRA results with 95% confidence intervals on the Aria glasses recordings according to the number of microphones and location.

4.2 Audio Quality Subjective Tests

Subjective listening tests on the EgoMusic were also conducted using a Multi-Stimulus Hidden Reference and Anchor (MUSHRA) test [1] through a web-based platform called GoListen [3]. In this test, participants were asked to grade on a 100-point scale the basic audio quality of a set of audio signals against a labelled reference signal. One trial contains: (1) a hidden reference signal identical to the labelled reference signal, (2) a hidden anchor signal which is a low-pass filtered reference signal with f_c = 3.5 kHz, and (3) four test signals to be evaluated. Two MUSHRA tests were conducted on two separate days to evaluate the recordings based on the number of microphones (MUSHRA-1) and location (MUSHRA-2).

For both MUSHRA tests, six trials corresponding to six distinct song snippets were presented to the participants. The six songs included two songs for each of sessions 1–3. These songs were selected based on their MOS from VISQOL, such that there was a large variation among the four test signals. For MUSHRA-1, the four test signals corresponded to the number of microphones (one, three, five, and seven) used for beamforming in Aria-near. For MUSHRA-2, the four test signals corresponded to mic1 from the four Aria glasses locations. The songs were presented in a randomized order for all participants. Twelve and eleven participants were recruited for MUSHRA-1 and MUSHRA-2, respectively. All participants rated themselves as having normal hearing. Participant ratings were only included when at least five out of six hidden reference signals were scored at least 80, and at least five out of six hidden anchor signals were scored less than 80. After post-screening, eight and seven participant ratings were obtained, respectively.

The MUSHRA test results are shown in Fig. 5. It can be observed that the subjective scores decrease with increasing distance from the sound source, similar to the objective results. However, the subjective scores decrease after applying beamforming, contrary to the objective results.

One-way ANOVA revealed that the subjective scores were significantly affected by beamforming (p = 0.01). Tukey-HSD further revealed that the mono group had significantly better scores compared to the 5-channel (p = 0.02) and 7-channel (p = 0.03) groups. Meanwhile, the subjective scores were not significantly affected by the location of the glasses (p = 0.08). To assess the agreement between the objective and subjective results, the Pearson correlation coefficient (R) was computed on the objective and subjective scores. Based on 24 data points, MUSHRA-1 had $R = -0.16$ while MUSHRA-2 had $R = 0.71$, signifying a strong linear relationship between the objective and subjective scores only when beamforming was not applied.

These audio quality tests suggest that the perceived audio quality captured by the Aria glasses is robust against its distance from the sound source within a performance theatre room (RQ1). Meanwhile, utilizing multiple microphones on the Aria glasses through a beamforming technique does not improve the perceived audio quality (RQ2). Moreover, the objective metrics used do not align with the subjective tests when beamforming was performed, signifying the need to develop objective metrics tailored for music performance capture through AAR.

4.3 Music Source Separation on the Data Set

A potential use case of AR glasses is to personalise the music listening experiences of individuals, especially for the hard-of-hearing (HoH), through a demixing-remixing approach [9, 10]. Thus, music source separation was employed on the Aria glasses recordings to determine the glasses' viability for personalised remixing. The *Hybrid Transformer Demucs* (*htdemucs*) model was used to separate the recordings into a *vocals*, *drums*, *bass*, and *other* stems [38].

Traditionally, music separation algorithms are benchmarked using the signal-to-distortion ratio (SDR) metric [49] on a standard dataset MUSDB18-HQ [33]. SDR calculates the ratio (in dB) between the reference source and the distortion artefacts produced in the separation. While the SDR can be computed for the clean recordings, it cannot be computed for the Aria glasses recordings since the reference sources for the latter were not obtained. Thus, the objective audio quality was evaluated using VISQOL for the estimated stems of the Aria glasses compared against the clean reference stems.

Table 3 tabulates the SDR results after using *htdemucs* on the test set of MUSDB18-HQ and the clean recordings of EgoMusic, where a higher SDR value indicates better separation quality. It was observed that *htdemucs* performs good stem separation on the 11 full-length songs of EgoMusic, except for the *drums* stem, since a different percussion instrument was used for EgoMusic compared

to the one used for training *htdemucs*. A two-stems evaluation was also conducted wherein the *drums*, *bass*, and *other* stems were mixed, yielding higher SDR results compared to the four-stems evaluation. EgoMusic obtained higher SDR results compared to the 50 full-length songs of MUSDB18-HQ because the latter had a larger and more diverse collection of music.

Table 3: The mean SDR for the stems of the data set

	4 stems				2 stems	
	vocals	drums	bass	other	vocals	other
MUSDB	8.10	9.77	8.03	4.36	8.10	14.49
EgoMusic	10.52	−6.09	12.95	6.98	10.52	15.06

Table 4 tabulates the average predicted MOS of the estimated stems of 55 audio samples. The estimated stems from the clean recordings yielded a higher MOS than that from the Aria recordings, which was expected. Furthermore, the estimated stems from the Aria recordings, excluding percussion, had higher MOS compared to their mixtures (see Table 2). Moreover, ANOVA revealed that the MOS was significantly affected by the Aria glasses location only on the *vocals* and *other* stems. Tukey-HSD revealed that the MOS of the *vocals* stem on Aria-static was significantly higher compared to Aria-near ($p = 0.001$) and Aria-far ($p = 0.003$). Meanwhile, the MOS of the *other* stems in the four-stems and the two-stems evaluation was significantly worse on Aria-mid compared to Aria-near ($p = 0.0002$, $p = 0.003$), Aria-far ($p = 0.0435$, $p = 0.012$), and Aria-static ($p = 0.0379$, $p = 0.002$).

These findings suggest that current MSS models provide satisfactory objective audio quality on the estimated stems (RQ3). Moreover, audio quality reduction with respect to distance becomes more pronounced for higher frequencies, i.e. *other* vs *bass*, indicating audio quality variation across the frequency spectrum. Subjective tests are recommended for further validation of these results.

Table 4: The average MOS on the stems and the p-value according to location.

	4 stems				2 stems	
	vocals	drums	bass	other	vocals	other
clean	4.08	1.87	4.56	3.93	4.08	4.09
near	3.20	1.08	4.46	3.09	3.20	3.22
mid	3.33	1.26	4.47	2.52	3.33	2.79
far	3.21	1.09	4.46	2.88	3.21	3.17
static	3.40	1.08	4.48	2.89	3.40	3.23
p_{loc}	**< 0.01**	0.42	0.77	**< 0.01**	**< 0.01**	**< 0.01**

5 Conclusion

This paper introduces EgoMusic, a novel egocentric multimodal dataset captured with Meta Aria glasses, addressing critical data scarcity for wearable Audio Augmented Reality (AAR) in music. Motivated by key challenges in audio fidelity, data capture, evaluation, and music processing on wearables, EgoMusic provides rich, synchronised multimodal recordings (7-channel audio, video, IMU, eye-tracking) with high-quality studio reference stems.

Our experimental findings aimed at understanding wearable capabilities for music processing and highlight several insights for the research community. Although current AR wearables such as the Aria glasses show fair audio quality (MUSHRA test) with robustness to listener distance (RQ1), achieving true musical fidelity remains challenging. Our study showed a gap between objective audio metrics (ViSQOL) and subjective human perception (MUSHRA), especially concerning beamforming for music. This suggests the need for a paradigm shift towards music-centric evaluation, prioritising the development of new perceptually validated objective measures for AAR music. Our results suggest that complex microphone array processing on current wearables does not guarantee superior perceived musical quality, calling for the investigation of simpler hardware for AAR music processing (RQ2). The promising source separation results using the htdemucs model confirmed the potential for personalised audio experiences (RQ3). Clipping was observed from the singing track of session 4 (musician perspective) due to the short distance between the source and the glasses' mics. Clipping demonstrates a crucial limitation of current AR wearables affecting musician-perspective processing.

Our findings and the EgoMusic dataset inform the community about the significant challenges of personalising music with wearables, representing a critical first step towards enabling impactful real-world AAR applications. Building on these insights, the multimodal nature of the EgoMusic dataset unlocks different research directions. It enables advancements in egocentric music processing, with researchers developing music-centric beamforming, noise reduction, and source localisation techniques benchmarked with EgoMusic to improve perceived quality on wearables. The EgoMusic rich data opens the exploration of multimodal music source separation, e.g., using audio-visual streams, eye-tracking for user attention, SLAM data for scene context, and IMU data for motion compensation to create more robust systems. The proposed dataset further supports the development of new perceptual metrics and a deeper understanding of listener attention in music AAR using contextual data such as eye gaze. Egocentric scene analysis also becomes viable, using SLAM for performance insights. This paper presents the first-ever music egocentric dataset for prototyping real-world AAR applications, such as on-device personalised remixing for hard-of-hearing listeners and context-aware music augmentation.

Acknowledgments

This work was conducted with research grants from Taighde Éireann – Research Ireland co-funded under the European Regional Development Fund under Grant Number 12/RC/2289_P2 and the European Commission Marie Skłodowska-Curie Actions grant ID 101151676. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. We thank the musicians and audience who participated in our data collection and Seán Clancy and Dunk Murphy for facilitating use of the the *Trapdoor* event space at the Creative Futures Academy, University College Dublin.

References

[1] 2015. *Method for the subjective assessment of intermediate quality level of audio systems.* Standard ITU-R BS.1534-3. ITU-R, Geneva.

[2] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Interspeech 2018*. 1561–1565. doi:10.21437/Interspeech.2018-1768

[3] Dan Barry, Qijian Zhang, Pheobe Wenyi Sun, and Andrew Hines. 2021. Go Listen: An End-to-End Online Listening Test Platform. *Journal of Open Research Software* (Jul 2021). doi:10.5334/jors.361

[4] Jonathan Berger, Ronald R Coifman, and Maxim J Goldberg. 1994. Removing noise from music using local trigonometric bases and wavelet packets. *Journal of the Audio Engineering Society* 42, 10 (1994), 808–818.

[5] P. Cabañas-Molero, A.J. Muñoz-Montoro, P. Vera-Candeas, and J. Ranilla. 2023. The music demixing machine: toward real-time remixing of classical music. *J Supercomput* 79 (2023), 14342–14357. doi:10.1007/s11227-023-05192-5

[6] Yunkee Chae, Junghyun Koo, Sungho Lee, and Kyogu Lee. 2023. Exploiting Time-Frequency Conformers for Music Audio Enhancement. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2362–2370.

[7] Thomas Chatzidimitris, Damianos Gavalas, and Despina Michael. 2016. SoundPacman: Audio augmented reality in location-based games. In *2016 18th Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 1–6.

[8] Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. 2020. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. arXiv:2004.09584 [eess.AS] https://arxiv.org/abs/2004.09584

[9] Gerardo Roa Dabike, Michael A. Akeroyd, Scott Bannister, Jon P. Barker, Trevor J. Cox, Bruno Fazenda, Jennifer Firth, Simone Graetzer, Alinka Greasley, Rebecca R. Vos, and William M. Whitmer. 2024. The first Cadenza challenges: using machine learning competitions to improve music for listeners with a hearing loss. (9 2024). http://arxiv.org/abs/2409.05095

[10] Gerardo Roa Dabike, Scott Bannister, Jennifer Firth, Simone Graetzer, Rebecca Vos, Michael A Akeroyd, Jon Barker, Trevor J Cox, Bruno Fazenda, Alinka Greasley, and William Whitmer. 2023. The First Cadenza Signal Processing Challenge: Improving Music for Those With a Hearing Loss. https://www.claritychallenge.org

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*.

[12] Shantanu Das, Seth Glickman, Fu Yen Hsiao, and Byunghwan Lee. 2017. Music everywhere–augmented reality piano improvisation learning system. In *Proc. International Conference on New Interfaces for Musical Expression (NIME)*. 511–512.

[13] Marta Sylvia Del Rio-Guerra, Jorge Martin-Gutierrez, Vicente A Lopez-Chao, Rodolfo Flores Parra, and Mario A Ramirez Sosa. 2019. AR graphic representation of musical notes for self-learning on guitar. *Applied Sciences* 9, 21 (2019), 4527.

[14] Emilie d'Olne, Alastair H Moore, Patrick A Naylor, Jacob Donley, Vladimir Tourbabin, and Thomas Lunner. 2023. Group conversations in Noisy environments (GiN)–Multimedia recordings for location-aware speech enhancement. *IEEE Open Journal of Signal Processing* (2023).

[15] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. 2021. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174* (2021).

[16] Callum Eaton and Hyunkook Lee. 2019. Quantifying factors of auditory immersion in virtual reality. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.

[17] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. 2023. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561* (2023).

[18] Giso Grimm, Joanna Luberadzka, and Volker Hohmann. 2019. A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta acustica united with acustica* 105, 3 (2019), 566–578.

[19] Pierre Guiraud, Sina Hafezi, Patrick A Naylor, Alastair H Moore, Jacob Donley, Vladimir Tourbabin, and Thomas Lunner. 2022. An introduction to the speech enhancement for augmented reality (spear) challenge. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 1–5.

[20] Soumya Shamarao Jahagirdar, Ajoy Mondal, Yuheng Ren, Omkar M Parkhi, and CV Jawahar. 2024. ICDAR 2024 Competition on Reading Documents Through Aria Glasses. In *International Conference on Document Analysis and Recognition*. Springer, 410–425.

[21] Nikhil Kandpal, Oriol Nieto, and Zeyu Jin. 2022. Music enhancement via image translation and vocoding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3124–3128.

[22] Joseph R Keebler, Travis J Wiltshire, Dustin C Smith, Stephen M Fiore, and Jeffrey S Bedwell. 2014. Shifting the paradigm of music instruction: implications of embodiment stemming from an augmented reality guitar learning system. *Frontiers in psychology* 5 (2014), 471.

[23] Shalva Kohen, Carmine Elvezio, and Steven Feiner. 2020. Mixr: A hybrid AR sheet music interface for live performance. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 76–77.

[24] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. 2025. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*. Springer, 192–210.

[25] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. 2024. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349* (2024).

[26] Eloi Moliner and Vesa Välimäki. 2022. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 943–956.

[27] Luc Nijs and Bahareh Behzadaval. 2024. Laying the foundation for augmented reality in music education. *IEEE Access* (2024).

[28] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. 2020. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[29] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. 2023. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20133–20143.

[30] Natasa Paterson, Katsiaryna Naliuka, Soren Kristian Jensen, Tara Carrigy, Mads Haahr, and Fionnuala Conway. 2010. Design, implementation and evaluation of audio for a location aware augmented reality game. In *Proceedings of the 3rd International Conference on Fun and Games*. 149–156.

[31] Nenad Petrović. 2020. Augmented and virtual reality web applications for music stage performance. In *2020 55th international scientific conference on information, communication and energy systems and technologies (ICEST)*. IEEE, 33–36.

[32] J. Pons, J. Janer, T. Rode, and W. Nogueira. 2016. Remixing music using source separation algorithms to improve the musical experience of cochlear implant users. *J. Acoust. Soc. Am.* 140(6) (2016), 4338–4349. doi:10.1121/1.4968597

[33] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. MUSDB18-HQ - an uncompressed version of MUSDB18. doi:10.5281/zenodo.3338373

[34] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. 2022. Automatic quality assessment of digitized and restored sound archives. *Journal of the Audio Engineering Society* (2022).

[35] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. 2024. Open-Source Conversational AI with SpeechBrain 1.0. arXiv:2407.00463 [cs.LG] https://arxiv.org/abs/2407.00463

[36] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624 [eess.AS] arXiv:2106.04624.

[37] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. 2021. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding* 211 (2021), 103252.

[38] Simon Rouard, Francisco Massa, and Alexandre Défossez. 2022. Hybrid Transformers for Music Source Separation. (11 2022). http://arxiv.org/abs/2211.08553 - Hybrid: time and spectral domain- self-attention within one domain, cross-attention across domains.- basis was wave u-net- made of two u-nets, one in time, and one in frequency- output of spectral branch is transformed to waveform using iSTFT, and summed with temporal branch, giving the actual prediction..

[39] Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2020. DiPCo — Dinner Party Corpus. In *Interspeech 2020*. 434–436. doi:10.21437/Interspeech.2020-2800

[40] Kouhei Sekiguchi, Aditya Arie Nugraha, Yicheng Du, Yoshiaki Bando, Mathieu Fontaine, and Kazuyoshi Yoshii. 2022. Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9266–9273.

[41] Donghee Shin. 2019. How does immersion work in augmented reality games? A user-centric view of immersion and engagement. *Information, Communication & Society* 22, 9 (2019), 1212–1229.

[42] Colm Sloan, Naomi Harte, Damien Kelly, Anil C Kokaram, and Andrew Hines. 2017. Objective assessment of perceptual audio quality using ViSQOLAudio.

IEEE Transactions on Broadcasting 63, 4 (2017), 693–705.

[43] Benjamin Stahl and Alois Sontacchi. 2023. Multichannel subband-fullband gated convolutional recurrent neural network for direction-based speech enhancement with head-mounted microphone arrays. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 1–5.

[44] Alex Stupakov, Evan Hanusa, Jeff Bilmes, and Dieter Fox. 2009. COSINE-a corpus of multi-party conversational speech in noisy environments. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4153–4156.

[45] S. Tahmasebi, T. Gajęcki, and W. Nogueira. 2020. Design and evaluation of a real-time audio source separation algorithm to remix music for cochlear implant users. *Front. Neurosci.* 14 (2020), 434. doi:10.3389/fnins.2020.00434

[46] Luca Turchet, Rob Hamilton, and Anil Çamci. 2021. Music in extended realities. *IEEE Access* 9 (2021), 15810–15832.

[47] European Broadcast Union. 2023. EBU R128 Loudness Normalisation and Permitted Maximum Level of Audio Signals.

[48] Vesa Valimaki, Andreas Franck, Jussi Ramo, Hannes Gamper, and Lauri Savioja. 2015. Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments. *IEEE Signal Processing Magazine* 32, 2 (2015), 92–99.

[49] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. 2006. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing* 14 (7 2006), 1462–1469. Issue 4. doi:10.1109/TSA.2005.858005

[50] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. 2020. CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings. In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*. 1–7. doi:10.21437/CHiME.2020-1

[51] Hagen Wierstorf, Dominic Ward, Russell Mason, Emad M Grais, Chris Hummersone, and Mark D Plumbley. 2017. Perceptual evaluation of source separation for remixing music. In *Audio Engineering Society Convention 143*. Audio Engineering Society.

[52] Xiao Xiao, Basheer Tome, and Hiroshi Ishii. 2014. Andante: Walking Figures on the Piano Keyboard to Visualize Musical Motion.. In *NIME*. Cambridge, MA, 629–632.

[53] Haici Yang, Shivani Firodiya, Nicholas J. Bryan, and Minje Kim. 2022. Don't Separate, Learn To Remix: End-To-End Neural Remixing With Joint Optimization. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 116–120. doi:10.1109/ICASSP43922.2022.9746077

[54] Yufeng Yang, Desh Raj, Ju Lin, Niko Moritz, Junteng Jia, Gil Keren, Egor Lakomkin, Yiteng Huang, Jacob Donley, Jay Mahadeokar, et al. 2024. M-best-rq: A multi-channel speech foundation model for smart glasses. *arXiv preprint arXiv:2409.11494* (2024).

[55] Jinzheng Zhao, Yong Xu, Xinyuan Qian, and Wenwu Wang. 2023. Audio Visual Speaker Localization from EgoCentric Views. *arXiv preprint arXiv:2309.16308* (2023).