IEEE Open Journal of Signal Processing

Received 22 November 2024; revised 10 April 2025; accepted 31 May 2025. Date of publication 10 June 2025; date of current version 25 June 2025. The review of this article was arranged by Associate Editor Z. Duan.

Digital Object Identifier 10.1109/OJSP.2025.3578299

The First Cadenza Challenges: Using Machine Learning Competitions to Improve Music for Listeners With a Hearing Loss

GERARDO ROA-DABIKE [1,2], MICHAEL A. AKEROYD [3], SCOTT BANNISTER [4], JON P. BARKER [2], TREVOR J. COX [1], BRUNO FAZENDA [1], JENNIFER FIRTH [3], SIMONE GRAETZER [1], ALINKA GREASLEY [4], REBECCA R. VOS [1], AND WILLIAM M. WHITMER [3]

[1]Acoustics Research Centre, University of Salford, M5 4WT Salford, U.K.
[2]School of Computer Science, University of Sheffield, Sheffield S10 2TN, U.K
[3]Hearing Sciences, School of Medicine, University of Nottingham, Nottingham NG7 2RD, U.K
[4]School of Music, University of Leeds, Leeds LS2 9JT, U.K

CORRESPONDING AUTHOR: GERARDO ROA-DABIKE (e-mail: g.roadabike@sheffield.ac.uk).

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) through Cadenza Project under Grant EP/W019434/1.

ABSTRACT Listening to music can be an issue for those with a hearing impairment, and hearing aids are not a universal solution. This paper details the first use of an open challenge methodology to improve the audio quality of music for those with hearing loss through machine learning. The first challenge (CAD1) had 9 participants. The second was a 2024 ICASSP grand challenge (ICASSP24), which attracted 17 entrants. The challenge tasks concerned demixing and remixing pop/rock music to allow a personalized rebalancing of the instruments in the mix, along with amplification to correct for raised hearing thresholds. The software baselines provided for entrants to build upon used two state-of-the-art demix algorithms: Hybrid Demucs and Open-Unmix. Objective evaluation used HAAQI, the Hearing-Aid Audio Quality Index. No entries improved on the best baseline in CAD1. It is suggested that this arose because demixing algorithms are relatively mature, and recent work has shown that access to large (private) datasets is needed to further improve performance. Learning from this, for ICASSP24 the scenario was made more difficult by using loudspeaker reproduction and specifying gains to be applied before remixing. This also made the scenario more useful for listening through hearing aids. Nine entrants scored better than the best ICASSP24 baseline. Most of the entrants used a refined version of Hybrid Demucs and NAL-R amplification. The highest scoring system combined the outputs of several demixing algorithms in an ensemble approach. These challenges are now open benchmarks for future research with freely available software and data.

INDEX TERMS Music, challenge, hearing aids, hearing impairment, hearing loss, machine learning.

I. INTRODUCTION

Most, if not all, human cultures have music [3]. Music brings people together, shapes society, and offers significant benefits to health and well-being [4]. Hearing loss can detract from the listening experience, however. The World Health Organization [5] estimates that by 2050 2.5 billion people will have some form of hearing loss, with at least 700 million requiring treatment. Hearing loss can lead to challenges with music such as: inaudibility of quieter passages, poor or anomalous pitch perception, and difficulty in identifying and distinguishing lyrics and instruments [6], [7], [8]. Therefore, it is essential to improve the processing of music in hearing aids and consumer devices, enabling those with hearing loss to continue enjoying and benefiting from music.

The most common intervention for mild to moderately severe hearing loss is hearing aids. Many of these devices have music programs but efficacy is mixed [9], [10], [11], [12]. For example, Greasley et al. [9] found that 68% of users report difficulties listening to music using their hearing aids. The issue is complicated because hearing aids are

typically frequency-dependent, nonlinear amplifiers to compensate for an individual's elevated hearing thresholds. In addition, they must allow for the rapid growth in loudness with low-intensity sound (loudness recruitment) and the potential discomfort from over-amplifying louder sounds. These wide dynamic range compression systems (WDRC) should make sound audible and comfortable. WDRCs alter the temporal envelope of the signal, however, with the degree of change dependent on how quickly they react to dynamic fluctuations. For example, they can introduce audible artifacts such as 'pumping'. Hearing aids also have features such as speech enhancement, feedback management, wind-noise reduction and scene analysis. The settings of hearing aids, from the frequency-dependent gain to how quickly the compressor reacts to additional features, are predominantly optimized for speech, and this means they may harm music, which has different spectral and temporal characteristics [10].

Research into hearing aid processing and music perception has indicated some approaches to improve audio quality, although the results are often mixed. Uys et al. [13] found that frequency compression that shifts the spectrum and/or envelope of high-frequency information in the signal to more audible lower frequencies, improved the self-reported quality across music genres for hearing-aid users with moderate to severe hearing loss. A later study found no statistically significant differences with frequency compression, however [14]. Croghan et al. [15] found that the quality of rock and classical music could be improved by using slow-acting rather than fast-acting WDRC. In contrast, Madsen et al. [16] found no significant overall effect of WDRC compression speed on a listeners' ability to hear individual instruments, although some participants found that slow-acting WDRC improved subjective clarity. These studies were based around traditional signal processing approaches, however, and nowadays, machine learning is the dominant paradigm in new audio processing algorithms. While machine-learning techniques have shown improvements in speech intelligibility for hearing-aid algorithms, e.g. [17] and [18], there is a gap in knowledge about how machine learning can improve the perceived audio quality of music for those with a hearing loss.

The above work also did not consider what could be done beyond amplification and compression. Sound engineering approaches, such changing the balance between instruments, have potential to improve audio quality for those with a hearing loss. Benjamin and Siedenburg [19] explored how listener preference was changed for pop music by altering lead-to-accompaniment level ratio, low-to-high frequency spectral balance and transformed equalization – i.e. equalization applied in a transformed domain, such as the power spectrum, rather than directly in the time or frequency domain. Elevated lead-to-accompaniment level ratio and music that was spectrally sparser was preferred by those with hearing loss.

In signal processing, many advances have been driven by open machine learning challenges (competitions), e.g. [20], [21], [22]. By providing a challenge infrastructure, including open databases for machine learning and specialized software tools, challenge organizers can significantly lower barriers that prevent out-of-field researchers from engaging in a topic. Challenges have also been shown to foster collaboration across disciplines, attracting a wider and more diverse range of researchers who contribute novel approaches to the field. Challenges also create a legacy through open benchmarks for future research. For the Cadenza project, both challenges were free to enter, with all the materials being provided at no cost to encourage as many entrants as possible.

The first application of a challenge methodology to the problem of improving audio quality of music for listeners with hearing loss is presented below. Two challenges are reported, the primary difference being that the first Cadenza Challenge (CAD1) [23] from 2023 was for listening over headphones, and the second, the 2024 ICASSP Grand Challenge (ICASSP24) [24], was for listening over loudspeakers. The tasks targeted demixing of stereo music signals followed by remixing. This was done because it allows sound engineering approaches to address issues. For example, people with hearing loss can struggle with lyric intelligibility [9], and this might be addressed by amplify the vocals between demix and remix [19]. Demixing was also chosen because there was an existing research community to tap into. Previous challenges in demixing are the Signal Separation Evaluation Campaigns (SiSEC) 2015-18 [25], [26], [27] and the Music Demixing Challenges MDX2021 and SDX2023 [22], [28], although none considered listeners with hearing loss. Building on these previous challenges, the demixing was into vocal, drums, bass and other instrument stems (VDBO). In both challenges, the objective metric was the Hearing Aid Audio Quality Index (HAAQI) [29].

The paper details the materials and methods developed for the two challenges, outlining the reasoning behind the scenarios, rules, baselines and data. This is followed by an evaluation and analysis of the entries. The paper finishes with a critique of the challenges and how this is informing future work.

II. MATERIAL AND METHODS

A. OVERVIEW

For CAD1, the scenario was listening to music over headphones without hearing aids, and entrants were given the left and right headphone input signals to process - see Fig. 1. For ICASSP24, music was reproduced via stereo loudspeakers with listeners wearing hearing aids. Thus, the left and right signals to be processed by entrants were from the hearing aid microphones. In both challenges, entrants were were asked to create a system that could rebalance the levels of the vocal, drums, bass and other instruments (VDBO). This would then allow for personalized mixes for people with a hearing loss. The VDBO representation was chosen because of its use in previous demixing research.

Fig. 2 shows the general structure of the challenge. Entrants were presented with scenes (blue box) containing a music extract to process and metadata giving the rendering requirements for the sample. For example, in ICASSP24 the metadata

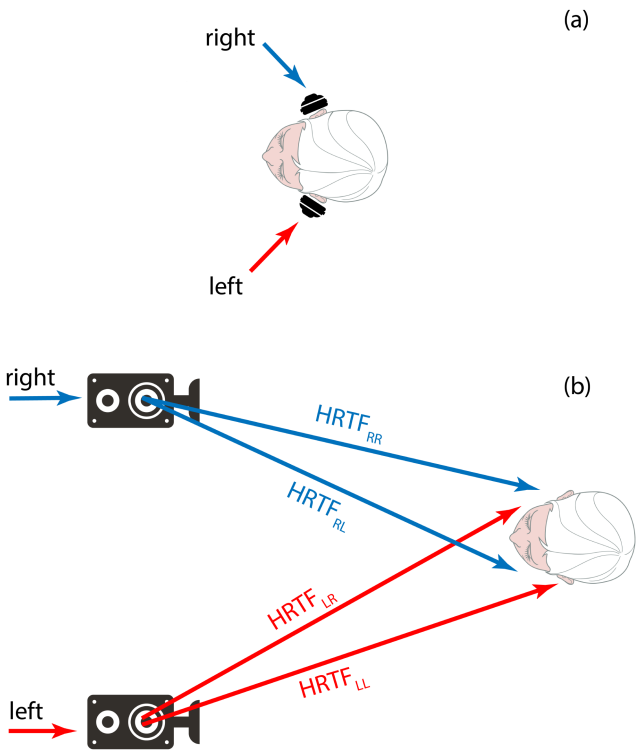

FIGURE 1. The scenarios for (a) CAD1 headphone listening and (b) ICASSP24 loudspeaker listening via hearing aids. HRTF, Head-Related Transfer Function.

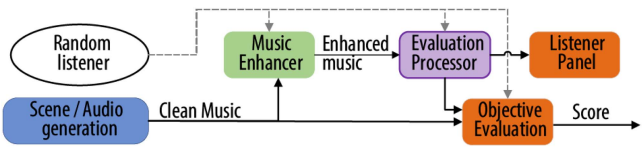

FIGURE 2. General structure of the challenges.

specified gains that should be applied to the VDBO signals before remixing. Additionally, a random listener (white oval) was selected from a database using a uniform probability distribution; this gave a pair of left and right ear audiograms to allow personalization of the signal processing and evaluation. It was hypothesized that hearing loss could be allowed for in the source separation algorithms. For instance, ensuring that separation artifacts were below the elevated hearing thresholds. A challenge rule specified that entrants were only allowed to modify the Music Enhancer (green box). The Evaluation Processor (lilac box) prepared the samples for either objective evaluation using HAAQI or perceptual testing via listeners with hearing loss.

Table 1 compares the CAD1 and ICASSP24 challenges. In CAD1, the music to be processed was the stereo signals being fed to a pair of headphones. In contrast for ICASSP24, the music came from left and right hearing aid signals when listening over a stereo loudspeaker pair. This meant that for ICASSP24, the music to be processed was a mixture of the right and left loudspeaker signals – see Fig. 1(b). The sound propagation

TABLE 1. Differences Between CAD1 and ICASSP24 Challenges

	CAD1	ICASSP24
Listening via	Headphones	Loudspeakers
Gains applied before remixing	No	Yes
Evaluation	HAAQI and listening panel	HAAQI

from the loudspeakers to the hearing aid microphones were modeled using head-related transfer functions (HRTFs). How the left and right signals from the loudspeakers combine at the ears is frequency dependent due to diffraction, reflection and interference around the head and hearing aids. This created additional complexities for ICASSP24 systems, compared to CAD1 and previous demixing challenges.

For CAD1, there was both objective and perceptual evaluation. Whereas, for ICASSP24 only objective evaluation was done because of time contraints in ICASSP Grand Challenges. In this paper, space-constraints mean only the objective evaluation for the two challenges are given. The CAD1 listening panel experimental design and results will be presented in a companion paper.

A final difference between the two challenges was that in CAD1 the separated VDBO signals were simply remixed back to stereo, whereas in ICASSP24 there were specified gains to be applied to the VDBO signals before remixing. The gains were added to test whether the systems submitted to ICASSP24 were capable of rebalancing the mix. Changing the levels between the VDBO components also makes artifacts created in the processing less likely to be masked and highlight cases where separation is imperfect. For example, in CAD1 if some of the drums was wrongly put into the bass track, then when the VDBO were summed together to give the stereo remix, the demix failure would be hidden. In ICASSP24, when there were different gains for the drums and bass tracks, this demix failure would result in the stereo remix being wrong.

CAD1 and ICASSP24 challenges did not have monetary prizes. CAD1 participants were invited to present their work in an online workshop. For ICASSP24, the top five entrants were invited to submit a two-page paper to the ICASSP 2024 conference.

B. OVERVIEW OF DATABASES

The data were split into training, validation and evaluation sets. Both the training and validation datasets were provided when the challenge launched, and were used by the teams to develop their signal processing systems. Training data is used to update the machine learning algorithm, whereas validation data is used to monitor the progress of the training and check for overfitting. The evaluation set tested generalization of systems to different music and listeners. The evaluation data did

not include the ground-truth, and was only made available a few weeks before the submission deadline. The challenge rules stated that teams should not use the evaluation data to improve their system.

Having access to large private datasets can give teams an unfair advantage when applying machine learning. For this reason, the challenge rules specified that teams could only use datasets and pre-trained models available in the public domain. However, entrants were allowed to augment the data using simple processing to create more robust systems. For example, they could randomize the VDBO stems, flip the right and left channels, apply SpecAugmentation [30] – a technique involving feature warping, frequency channel masking, or time-step masking – and pitch shifting.

C. LISTENER DATABASES

Each music extract needed to be personalized to allow for the hearing acuity of a target listener. The hearing was characterized by bilateral pure-tone audiograms at the standardized frequencies of 250, 500, 1000, 2000, 3000, 4000, 6000, 8000 Hz. The bandwidth of music is wider than this, but we were limited by the available datasets of audiograms. The data were anonymous audiograms from bilateral hearing aid users. Hearing loss levels at each frequency were limited to 80 dB to be consistent with the training dataset from the Clarity Project [31] that we were using. This limit was applied in Clarity because (i) the hearing loss model they used produces unrealistic signals for large impairments, and (ii) the headsets they used in listening tests could not reproduce high enough levels to compensate for large impairments.

Hearing loss severities were based on the mean, better-ear 4-frequency (500, 1000, 2000 and 4000 Hz) hearing loss criteria [32]. These were no loss (0–19 dB), mild (20–34 dB), moderate (35–49 dB), moderately severe (50–64 dB), severe (65–79 dB) and profound (>= 80dB). The datasets were as follows:

- Training: 83 audiograms from the Clarity Project [31]. These correspond to real, anonymized audiograms drawn from the participant database of the Scottish Section of Hearing Sciences at the University of Nottingham. There were no people with no loss, 17 people with mild, 44 with moderate, 22 with moderately severe, and none with severe.
- Validation: 50 audiograms drawn from von Gablenz et al. [33]. The audiograms were randomly selected to have the same distribution as the training set. First, audiograms were filtered using the better-ear 4-frequency hearing loss criteria, with thresholds between 20 and 75 dB. Then, the audiograms were randomly chosen to maintain the same distribution per band as in the training set. This set had an equal male-female distribution. The distribution was no people with no loss, 24 with mild, 22 with moderate, 4 with moderately severe and 0 with severe.
- Evaluation: 53 audiograms; 52 listeners with a hearing loss were recruited for the Cadenza listening panel by the University of Leeds, U.K. An additional listener with an audiogram with 0 dB loss at all frequencies was included. The distribution was 3 listeners with no loss, 13 with mild, 17 with moderate, 19 with moderately severe and 1 with severe.

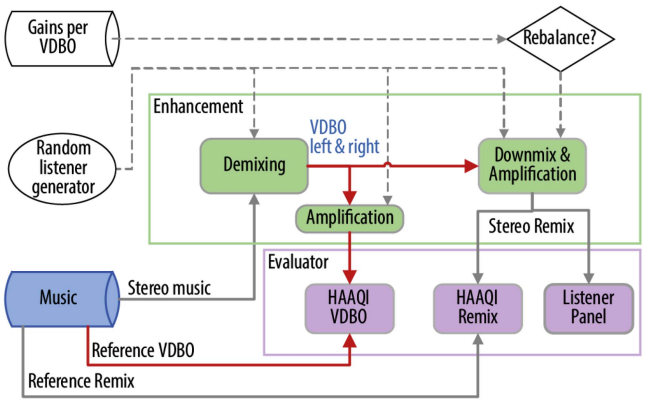

FIGURE 3. Baseline architecture for CAD1 and ICASSP24. The gains were only applied in ICASSP24.

D. MUSIC DATABASES

There are established public datasets that are benchmarks for demixing challenges. These were used for CAD1 and ICASSP24 for comparability with previous work. The music for training, validation and evaluation used the standard splits for MUSDB18-HQ [34], giving 86, 14 and 50 stereo tracks respectively. MUSDB18-HQ contains isolated stems for vocals, drums, bass and other (VDBO), as well as stereo mixes. The music is mostly Western pop/rock with a small amount of reggae, rap, heavy metal, and electronic.

An independent validation set was constructed by randomly selecting 50 tracks from the MoisesDB dataset [35], while maintaining the same genre distribution as the evaluation split of MUSDB18-HQ. This new validation set was included because many pretrained models that use MUSDB18-HQ, incorporate the MUSDB18-HQ validation split as part of their training.

E. BASELINE SOFTWARE TOOLS

The baseline is a complete software system that can run the task. It includes a solution to the problem in the music enhancer. Fig. 3 shows the architecture. The problem was presented as a demix/remix task, with a view to allowing listeners to rebalance and personalize a mix. Systems needed to take stereo pop/rock music and demix it into VBDO signals. In ICASSP24, gains would then be applied to these four signals before they were remixed back to stereo. This is similar to previous demix challenges [22], [27], [28], except that entrants could allow each listener's hearing loss in the demixing. An additional novelty compared to previous demix challenges, was that the evaluation metric was tailored for listeners with hearing loss and hearing aids. HAAQI, the Hearing-Aid Audio

Quality Index was used [29]. While the baseline did demixing, the challenge rules allowed entrants to submit stereo audio from an end-to-end system without an explicit demixing stage. However, no entrant chose to do this.

Two baseline demix algorithms were given to entrants. These were out-of-the-box pretrained audio source separation algorithms with no retraining. One used the Hybrid Demucs model [36] (HDemucs), which employs a U-Net architecture to combine both time-domain and spectrogram-based audio source separation. The other used the Open-Unmix model [37], which just uses spectrograms. HDemucs is probably the most commonly used demixing algorithm, whereas Open-Unmix is simpler to implement and train.

The music enhancer also needed a frequency-dependent amplification stage to correct for the raised auditory thresholds due to hearing loss. NAL-R [38] was used to match the default amplification applied to the reference signal during the HAAQI evaluation (see below). As HAAQI compares the processed signal with a reference, the frequency-dependent amplification stage needs to be the same for the two to maximize HAAQI scores.

F. LATENCY, CAUSALITY AND MODEL SIZE

For a signal processing system to be used with live music on a hearing aid, it needs to operate with low latency. This restricts what machine learning approaches can be used. However, listening to recorded music is also very common, so there are scenarios where latency is not a concern. For these reasons, the rules allowed both causal and non-causal entries. HDemucs and Open-Unmix are non-causal demixing models, which means they have access to future samples to inform processing. In contrast, causal models rely only on current and past information, and this harms performance. This can be illustrated using Convolutional TasNet models applied to source separation, as causal and non-causal versions are available [39]. The casual version has better performance for speaker separation, achieving 10.6 dB signal-to-distortion ratio (SDR), compared to it's non-causal implementation, that has an SDR of 15.6 dB SDR. Similarly, for classical music separation, the causal model achieves 5.10 dB SDR, whereas the non-causal model achieves 6.12 dB SDR.

For causal processing, the challenge rules restricted systems to only use input samples less than 5 ms in the future. The tolerability of delays can depend on numerous factors, from the type of stimulus (e.g. speech vs. music), availability of visual cues as well as different aspects of the impairment – gain based on severity of loss, loudness recruitment and temporal fine structure resolution [40]. Numerous studies have shown deleterious effects of delay beginning as early as 5 ms (e.g. [40], [41]), hence this delay was chosen as the upper limit in the Cadenza rules.

Hearing aids cannot currently host the huge deep neural networks that are common in audio processing. However, the challenges did not place a limit on the model size or computing resources being used by systems. The reason for this is that a common route to innovation is to first produce solutions that are computationally expensive and then apply methods such as knowledge distillation [42] to reduce the resources required while maintaining performance.

G. RENDERING AND REMIX METADATA FOR ICASSP24

For the ICASSP24 challenge, the scene generator had to simulate loudspeaker reproduction. This was done by applying Head Related Transfer Functions (HRTFs) measured in anechoic conditions - see Fig. 1(b). The scene generator randomly selected one of the 16 measured human heads from the OlHeaD-HRTF dataset [43], the azimuth angle of the listener's head, and then extracted the appropriate HRTFs from the dataset. These HRTFs were for Behind-The-Ear (BTE) hearing aids with three microphones on each side to allow for beam forming to improve signal to noise ratio for speech in noise. But for the Cadenza music challenges just the front left and right microphones of the hearing aids were used.

Listeners were modeled to have a variety of head orientations around the azimuth range for standard stereo loudspeaker reproduction (i.e., around ±30°). This was to simulate non-perfect stereo reproduction. Angles were all nine combinations from ±22.5°, ±30° and ±37.5°.

Each music track was divided into several consecutive 10-second excerpts, ensuring that no silent portions were selected. Then an HRTF pair was applied to each excerpt. This means that two excerpts from the same track had different pairs of HRTFs applied, thus requiring separation models to be robust under varying HRTF conditions and for different songs.

For ICASSP24, the generator also randomly set the gains to be applied to each VDBO stem before the remix to stereo. Unfortunately, there was little prior knowledge to guide what might be the preferred gains, and furthermore these would vary with listener preference and the music. Consequently, random gains were used to bracket significant changes in the stereo remix to create systems that could enable any remix a listener might ask for. First, the number of VDBO stems that had their gain altered was randomly chosen using a uniform probability distribution (i.e., 1, 2 or 3). Then the gains were chosen for each of these tracks from ±10 dB, ±6 dB and ±3 dB. Again, a uniform probability random distribution was used.

H. EVALUATION

The objective evaluation was done using HAAQI [29]. HAAQI was used as it is the only published metric for audio quality of music for listeners with a hearing loss listening through hearing aids. HAAQI was developed based on perceptual ratings from 34 listeners auditioning three music excepts that passed through 100 different signal processes. One limitation is that the signal processes used to develop HAAQI will differ from those created by machine learning. However, the use of the relatively sophisticated auditory model within the calculation of HAAQI adds robustness compared to a purely empirical model. Nevertheless, HAAQI is only based

on thresholds and does not consider supra-threshold information or the effects of the age of the listener. As with all objective metrics, HAAQI can only approximate perception. For example, it is a monaural measure and therefore does not model effects such as binaural unmasking. In the challenge, the metric was an average of the HAAQI scores calculated for the left and right ears separately. A final issue, is that the music samples used during the development of HAAQI did not include pop/rock. However, others have subsequently used it on different genres. One study on adaptive feedback cancellation algorithms found a correlation coefficient of 0.81 between HAAQI scores and sound quality ratings for 8 pieces of music [44]. Another study into codecs found correlation coefficients between 0.68 and 0.92 for HAAQI and perceptual ratings for solo and ensemble music with various genres [45]. A final study found a correlation coefficient of 0.68 between HAAQI and audio quality ratings [46] for genres including pop/rock. These three studies only considered listeners with normal hearing, however.

HAAQI is an intrusive metric that requires a reference signal to assess the audio quality of a processed signal. HAAQI incorporates an auditory model that accounts for hearing loss through the listener's audiogram and prescriptive gain. It evaluates the temporal fine structure (basilar membrane vibration) and the envelope of the reference and processed signals, measuring differences via correlation and spectral analysis. HAAQI combines linear and nonlinear terms to predict perceived quality. HAAQI effectively predicts quality changes due to additive noise, nonlinearities, and spectral shifts.

For the VDBO evaluation, HAAQI was calculated for the 8 stems (VDBO signals for left and right sides) and then an average taken. The reference signals corresponded to the ground-truth VDBO. For evaluating the remix stereo, HAAQI was calculated for the left and right signals, and an average taken. For the CAD1 remix, the ground-truth mixtures were used as the reference signal. For the ICASSP24 remix, the reference signal corresponded to a remix of the ground-truth VDBO signals using the specified HRTF and gains.

HAAQI was developed as a perceptual model and so is relatively slow to compute and non-differentiable. These factors make it hard to incorporate into machine learning efficiently. Furthermore, it is an intrusive metric and therefore requires a reference signal. This reference needs a frequency-dependent amplification to correct for raised hearing thresholds. Whatever amplification scheme is chosen must be replicated in the music enhancer, otherwise the HAAQI value decreases. In both CAD1 and ICASSP24, a linear FIR filter was used based on the NAL-R prescription method [47], based on a public-domain implementation. While most hearing aids use dynamic range compressors operating over different frequency ranges, the best settings for these compressors are disputed (see Introduction).

For CAD1, 49 out of the 50 music tracks in the MUSDB18 -HQ evaluation split were used. Because of the subjective evaluation by listeners, one track was excluded due to offensive words in the lyrics. To keep the submission size within reasonable bounds (around 23 GB; 4 VDBOs for the left channel, 4 VDBOs for the right channel, and 1 remix for each listener), entrants were required to submit 30-second extracts for the VDBO and 15-second extracts for the remixed. These extracts were selected randomly, ensuring that all VDBO stems were active at some point. Each extract was processed for all 53 listeners, obtaining N = 2,597 processed extracts per system. For ICASSP24, all 50 evaluation tracks from MUSDB18-HQ were used. The tracks were segmented into consecutive 10-second extracts, resulting in 960 audio segments. For each of these 960 extracts, the music was paired with a random HRTF and a random gain. To keep the submission package around 20 GB, each of these were processed for 20 random listeners from the pool of 53 listeners, giving N = 19,200 audio examples tested per system.

III. RESULTS

A. CAD1 CHALLENGE

Seven entries, two baselines and a do-nothing system were evaluated. Table 2 summarizes the different approaches of the systems for CAD1 and the average HAAQI scores.

Eight systems used either the HDemucs or OpenUnmix models for source separation. One of those (E05) refined OpenUnmix by using a sliced Constant-Q Transform (sliCQT) [49] with the Bark scale; a neural network architecture that used a convolutional denoising autoencoder (CDAE) [49], [50], and all targets were trained together with combined loss functions like CrossNet-Open-Unmix (X-UMX) [51].

Six systems did not alter the remix from the baseline, meaning the demixed VDBO stems were simply added together to get the stereo output. Those who did alter the balance between the VDBO stems included E17. This system applied a mid-side equalization, a technique that separates the stereo signal into mid (center) and side (stereo width) components. By independently processing the mid and side signals, one can enhance the tonal balance of centrally placed elements, such as vocals or bass, while also adjusting the stereo width to improve spatial characteristics. In the mid-side equalization, the new left signal L' and new right signal R' are given by:

$$\begin{aligned} L' &= G(M) + H(S) \\ R' &= G(M) - H(S) \end{aligned} \tag{1}$$

where M is the mid signal and S the side signal given by:

$$\begin{aligned} M &= (L + R)/2 \\ S &= (L - R)/2 \end{aligned} \tag{2}$$

The function *G()* was two parallel filters that reduced the mid signal below 2 kHz by 2 dB, to attenuate frequencies that were not part of the lead vocals. The function *H()* was three parallel filters to increase the side signal between 2 and 6 kHz by 3 dB to help with binaural unmasking. E17 also applied a single compressor to the amplified remixed signal.

TABLE 2. Overview of System Approaches and Scores for the CAD1 Challenge.

System	Separation	Remix	Amplification	HAAQI VDBO	HAAQI remix
E01, Baseline 1	HDemucs	Original	NAL-R	**0.255 ± 0.041**	**0.706 ± 0.196**
E02, Baseline 2	OpenUnmix	Original	NAL-R	0.225 ± 0.029	0.638 ± 0.161
E05	OpenUnmix*	Original	NAL-R	0.094 ± 0.014	0.677 ± 0.186
E12	HDemucs	Rebalanced	Multiband compressor	**0.255 ± 0.041**	0.684 ± 0.205
E14	HDemucs	Original	NAL-R*	0.203 ± 0.029	0.530 ± 0.226
E15	OpenUnmix	Original	NAL-R*	0.183 ± 0.023	0.475 ± 0.191
E21	-	Original	None	0.421 ± 0.216	0.440 ± 0.234
E16	Spleeter [48]	Original	NAL-R*	0.135 ± 0.027	0.270 ± 0.148
E17	HDemucs	Mid-Side EQ	NAL-R + compressor	0.236 ± 0.033	0.276 ± 0.105
E22	HDemucs	Rebalanced	NAL-R	0.195 ± 0.039	0.217 ± 0.109

Total 10 systems with 2,597 processed audios per system. "Original" remix means using the gains specified for the original music. * indicates a refined version of the algorithm used. HAAQI scores are averaged over the evaluation set with ± standard deviations, with VDBO being for the separated VDBO stems and remix for the output stereo. Finally, average HAAQI results across the evaluation set are shown with standard deviations as ± values (N=2,597). Values in bold show the highest scores. E21 is a do nothing system where the processed signals were equal to the original signals with no amplification.

Another two systems that changed the balance in the remix were E22 and E12. Both used methods to increase the prominence of the vocal track. For example, E22 used gains of +7.6, −8.0, −4.4, and −4.4 dB for the VDBO stems when all were not silent. E12 decreased the level of the non-vocal tracks for people with moderate or severe hearing loss. Since the Other stem may include instruments like guitar and piano, E12 used a multiband compressor to reduce the dynamics of the Other stem, giving more space for the vocals. The compression thresholds were set based on the levels of the vocals.

Changes to the EQ or VDBO balance would decrease HAAQI scores. This arises because HAAQI is intrusive and compares the processed signal to a reference. HAAQI is an approximate model of perception, however, and changes to EQ or balance could improve scores in the listening tests.

Four systems used the NAL-R amplification provided in the baseline. The exceptions were: (1) E12, which used a multiband compressor. (2) E14 and E15, which used a linear filter like NAL-R but decreased the low-frequency attenuation relative to the original NAL-R algorithm. The 250 and 500 Hz bands were increased by 16 and 7 dB, respectively. This was intended to increase the bass, but it had the unintentional consequence of limiting the high-frequency amplification where hearing loss is usually most significant. This limitation arose because the broadband signals were peak normalized in the time domain to prevent clipping. (3) E16 applied a Butterworth bandpass filter with −3 dB points at 250 Hz and 18.5 kHz. For all of these systems, the departure from NAL-R decreases the HAAQI score but has potential to improve listening test scores.

The HAAQI scores averaged for the VDBO stems in Table 2. These need to be read with some caution because applying HAAQI to individual stems is untested. The objective scores for the VDBO signals show that Baseline 1, which used HDemucs, scored highest. In setting up the challenge, it was thought that allowing for hearing loss in the demix processing might improve performance. For instance, artifacts created during source separation might fall below the elevated hearing threshold. But no system exploited this possibility. There have been many demixing challenges, which means that the state-of-the-art approaches used in the baseline were hard to beat.

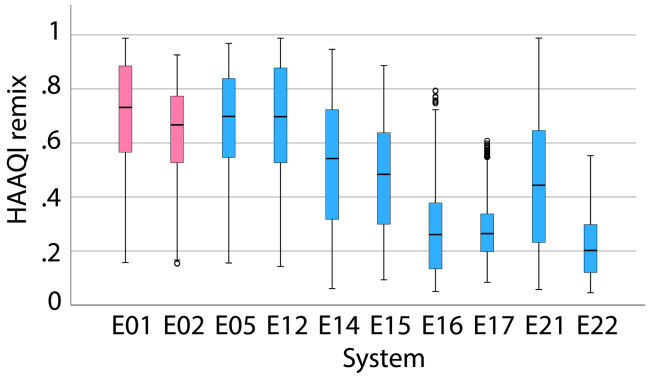

FIGURE 4. HAAQI scores for remix for each system for CAD1. Baseline systems are shown in pink.

The HAAQI scores for the remix stereo are shown in Table 2 and a box-plot is shown in Fig. 4. The data did not meet the assumptions needed to use an ANOVA. For example, the dependent samples were not drawn from a normally distributed population with evidence of a ceiling effect where HAAQI=1 for some teams. Consequently, the following is an analysis of main effects using non-parametric approaches.

A one-way Kruskal-Wallis test with HAAQI values as the dependent variable and systems as the independent variable was carried out. This showed that the differences between the scores for the ten systems were significant (N = 25,970, df=9, H=12,824, p < 0.001, η^2=0.49), with a very large effect size. Pairwise comparisons using the Mann-Whitney U test, showed that most systems were significantly different

from each other ($p < 0.001$ for pairs with significant difference, except E05-E01 where p=0.02. Bonferroni correction for multiple tests applied). The three pairs of systems with no significance difference were: E16 and E17 (p=1); E05 and E12 (p=1); and E12 and E01 (p=0.1).

Whether system performance varied with hearing loss severity was examined. The listener audiograms were coded into a 5-value ordinal variable: no loss, mild, moderate, moderately severe and severe. For this calculation, the average of the left and right audiograms was used rather than the better ear. This was done because the HAAQI metric were an average of the HAAQI values for the left and right signals. The Spearman rho between HAAQI and hearing loss severity was -0.540 ($N = 25{,}970$; $p < 0.001$). This means the HAAQI scores were lower for those with worse hearing loss, a trend seen for all systems. This trend explains 29% of the rank variance.

Overall, for the remix HAAQI scores, the HDemucs Baseline (E01) had the highest score. As noted above, the 5 systems that applied different remix or amplification systems were bound to score lower on HAAQI. These entrants used different amplification approaches to improve scores in the listening panel evaluation.

The lack of HAAQI improvement over the baseline, indicated a need for a scenario with more chance of bettering the baseline. This led to the ICASSP24 challenge. Specifically, the use of loudspeaker reproduction in ICASSP24 meant that out-of-the-box demix algorithms would perform worse because of the frequency-dependent mixing of the left and right music signals. Furthemore, the specified gains for the VDBO would highlight bleed between demixed components. This also motivated a push for causal systems because non-causal approaches would not work on hearing aids.

B. ICASSP24 CHALLENGE

There were 17 systems entered from 11 teams. Table 3 summarizes the average HAAQI scores for the different systems. Nearly all differences between the system scores in the table were statistically significant, but some had very small effect sizes (see later for statistical analysis). The table also summarizes the approaches for the different systems. Nine systems beat the best baseline, and the discussions below of the non-causal systems will concentrate on these.

The baselines were trained on the original stereo music and not on the hearing aid signals. Consequently, it was expected that retraining an established source separation system on the hearing aid signals would be sufficient to improve scores. Examples of teams doing this were T11 and T46.

The highest scoring system, T47, took an ensemble approach, with the output of the separation algorithm being an average of three systems. These were pretrained versions of Dual-Path TFC-TDF UNet [58], HDemucs, and a version of MDX-Net [59] only trained on the MUSDB18-HQ dataset. These were then fine tuned on the ICASSP24 dataset. T22 took a similar ensemble approach, but one of the two pretrained models used the label noise dataset from the Sound Demixing Challenge 2023 [22], which was outside the rules of the ICASSP24 challenge.

TABLE 3. Overview of the Approaches for ICASSP Challenge.

System	Separation	Amplification	HAAQI
Non-Causal			
T01, Baseline	HDemucs	NAL-R	0.570 ± 0.185
T02, Baseline	OpenUnmix	NAL-R	0.511 ± 0.153
T47 [52]	Ensemble	NAL-R	**0.632 ± 0.177**
T22	Ensemble	NAL-R	**0.631 ± 0.173**
T03-S [53]	HDemucs*	NAL-R	**0.593 ± 0.186**
T03 [53]	HDemucs*	NAL-R	**0.592 ± 0.185**
T11-A [54]	HDemucs*	NAL-R	**0.586 ± 0.188**
T18 [55]	U-Nets + DPRNN	?	**0.585 ± 0.183**
T11 [54]	HDemucs*	NAL-R	**0.580 ± 0.186**
T12	HT-Demucs [56]	?	**0.573 ± 0.182**
T46 [57]	HDemucs*	NAL-R	**0.570 ± 0.185**
T25	HDemucs	Compressor + NAL-R*	0.561 ± 0.163
T31-A	HT-Demucs*	?	0.543 ± 0.169
T42	HDemucs	?	0.543 ± 0.173
T42-A	HDemucs	?	0.534 ± 0.172
T31	HT-Demucs*	?	0.530 ± 0.174
T09-B	HDemucs	?	0.479 ± 0.132
T09	HDemucs	?	0.478 ± 0.135
Causal			
T16	k-means	?	0.144 ± 0.018

A: system with data augmentation. S: system with supplementary data. B: second submission. * indicates a refined version of the algorithm was used. ? indicates amplification model not specified in technical report and so assumed to be NAL-R from the baseline. Finally, average HAAQI results across the evaluation set. Standard deviation shown as ± value, N=19,200. Values in bold show systems that performed better than T01 Baseline.

There were some refinements for systems that built on established architectures. T03 and the version trained on supplementary data (T03-S) added 15% of the original stereo into the final mix using a skip connection that bypassed the demix/remix. The intention was to restore components that get lost in the demix/remix process. T11 introduced a modified logit function intended to create a larger gradient for hard-to-learn examples. This used self-knowledge distillation with progressive refinement of target (PS-KD) [60]. An ablation study showed that this modified logit produced a very small improvement in HAAQI of 0.009. T46 replaced the original complex ratio mask in HDemucs with a deep filter [57].

There were also some refinements on how the training data was used to improve learning. Audio data augmentation techniques achieved a very small improvement in HAAQI of 0.006 for T11-A vs T11. This team also explored curriculum learning where initial training used easier examples, before moving onto the harder cases after a set number of epochs. This produced only a very small improvement in HAAQI of 0.002, however.

Only T25 attempted to improve the amplification stage of the processing by applying a single compressor to the remixed signal and changing tuning the number of filter orders of NAL-R. As the reference signal in the objective evaluation involved amplification using NAL-R, this could only reduce the HAAQI scores. However, it is worth noting that NAL-R

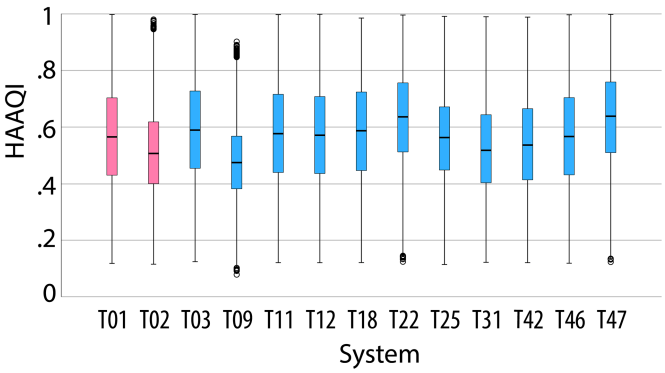

FIGURE 5. HAAQI scores vs system for ICASSP24. Baselines shown in pink.

does not account for all the non-linear level-dependencies typical with hearing loss, such as loudness recruitment, and so the approach of T25 might result in improved scores in listening tests, but that remains untested.

The causal system T16 did not score as well as the non-causal approaches. Previous demixing research has mostly focused on non-causal approaches, so there were preexisting, refined approaches to build upon for the Cadenza challenge. Furthermore, it would be expected that a causal method would score lower because any machine learning algorithm has less input information to work from compared to non-causal techniques. T16 used a *k*-means clustering based on 39-dimensional Mel-frequency cepstral coefficients (MFCC) features for the VDBO stems. MFCCs are features that estimate the shape of the smoothed spectrum to infer the configuration of the vocal tract. Then for the mixture, for every 5 ms frame, the system tried to identify which of the VDBO stems was dominant via the MFCC features, and then allocate the frame to the appropriate VDBO signal. However, such an approach struggles when more than one VDBO stem is prominent in a frame.

A statistical analysis of the ICASSP24 results was performed. First, T16 was removed as an outlier as its mean and standard deviation were both much smaller than for all other systems. The systems using supplementary data or data augmentation were also removed to ensure that scores from each system were statistically independent (e.g. T31 was analyzed but not T31-A). This left 13 systems. As with CAD1, the data did not meet the normality assumption needed to use an ANOVA, and therefore an analysis of main effects using nonparametric approaches was used.

Fig. 5 shows a box-plot of the HAAQI scores for each system. A one-way Kruskal-Wallis test with HAAQI as the dependent variable and system as the independent variable was significant ($N = 249{,}600$, $df = 12$, $H = 13{,}682$, $p < 0.001$, $\eta^2 = 0.05$). The effect size was small, however. Pairwise comparisons using the Mann-Whitney U test, showed that most systems were significantly different from each other ($p < 0.001$ for pairs with significant difference, except T25-T46 and T11-T18 where $p = 0.03$; Bonferroni correction for multiple tests was applied). The exceptions where there was no significant difference were: T01 and T12 ($p = 1$); T01 and T046 ($p = 1$); T01 and T25 ($p = 0.2$); T11 and T12 ($p = 0.2$); and T03 and T18 ($p = 0.2$).

It was hypothesized that the greater the differences in gains applied to the VDBO stems before remixing, the poorer the performance would be. The thinking here was that any bleed or artifacts created during demixing would be more evident in the remix. To test this, the HAAQI scores were correlated with the standard deviation of the gains applied to the VDBO stems. The Spearman's rho was −0.318 (N=249,600; $p < 0.001$). This indicates that HAAQI scores were indeed lower when there were larger differences in gains between the VDBO stems. This explained about 10% of the rank variance.

HAAQI scores were analyzed to determine whether they varied with hearing loss severity. This analysis used the same hearing severity classifications as for CAD1. The Spearman's rho between HAAQI and hearing loss severity was −0.454 (N=249,600; $p < 0.001$). This means that the more severe the hearing loss, the lower the HAAQI scores. This explains 20% of the rank variance. System T09 was the only one that did not have a linear relationship for hearing loss severity. For that system, the best scores were for moderate loss.

Greater hearing loss means that the NAL-R amplification would have been larger, especially at high frequencies. One possibility is that errors or artifacts in the demix are greater at higher frequencies, and hence the HAAQI score decreases for larger hearing loss. Another possible cause is the greater amplification created more clipping in some music extracts with larger hearing loss severity, which would then decrease the HAAQI score.

Exploring how the HAAQI scores varied with the angle between the listener and the left and right loudspeakers in the stereo reproduction yielded a result that was significant but had tiny effect sizes. Significance occurred because of the very large number of examples (N=249,600) but the differences in HAAQI scores were too small to be important.

IV. DISCUSSION

The Cadenza project created a series of machine-learning challenges to increase the number of music processing researchers considering hearing loss. We developed baseline software and curated open source datasets. The aim was to catalyze a cultural shift in the audio machine learning community, so more research includes the range of hearing abilities seen in the general population, rather than the default assumption of young 'normal hearing' [61].

Nowadays, one difficulty with using a challenge methodology is that the number of signal processing competitions makes it harder to get entrants. The increase in the number of entrants from CAD1 to ICASSP24 shows that Cadenza is beginning to grow the community. This has been achieved by engaging with researchers who work on music demixing. Using gatekeepers to raise awareness of challenges is helpful, which is why we ran a challenge as part of ICASSP. Continuing this, the next Cadenza Challenge, CAD2, is an official challenge of the IEEE Signal Processing Society. It will also have some modest cash prizes to encourage entrants.

Choosing appropriate tasks, rules and evaluation methods for a new series is difficult because what teams might be able to achieve can be hard to predict. In the CAD1 challenge, the baselines were based on state-of-the-art demixing models, which with the benefit of hindsight were difficult to beat. Learning from this, the ICASSP24 challenge made the problem more difficult by introducing loudspeaker reproduction and specified gains to be applied to the VDBO stems before the remix. The lower scores for the top systems in ICASSP24 compared to the best in CAD1, indicate that there is still scope for further research into the ICASSP24 scenario.

All but one entrant used non-causal signal processing, which means the methods could only be applied to recorded music or broadcast situations where a delay in processing is not an issue. For hearing aids and live music, low latency and causal methods are required. In future, more work is needed to encourage causal systems. Hence, the second Cadenza challenge CAD2, features a causal baseline. Future challenges may also consider limiting the size of the deep learning models, so they can be implemented at low latency on hearing devices.

The objective metrics currently available for machine learners need improving. HAAQI was used because it is the only audio quality metric that accounts for hearing loss and hearing aid processing. However, it is not ideal for machine learning because it is slow to compute and non-differentiable. Furthermore, as discussed above, the amplification stage used in HAAQI sets the gold-standard for entrants to try and achieve. In CAD1, some teams chose to create systems that deviated from NAL-R to improve scores in the listening tests, despite this reducing HAAQI. For this reason, CAD2 will move away from NAL-R to use non-linear amplification with parameter settings similar to those used in current hearing aids and the frequency-specific gains determined by each individual's pure-tone thresholds. Future work could go further, however, and consider how deep learning models might be used to overcome the limitations of classical signal processing of non-linear amplification.

A non-intrusive metric might overcome some of these issues. The audio created in CAD1 has been used in listening tests and work is ongoing to create a metric based on those results. For CAD2, one of the tasks is improving lyric intelligibility. For intelligibility assessment, we are using the Whisper model [62] to transcribe lyrics and compute word correct rates. Whisper does not require a reference signal.

Objective evaluation is always limited by the metric available, because this can only ever be an approximation to human listening. A true test of a system requires listening tests. These have been carried out on the audio submitted to the CAD1 challenge. The design of these experiments and results will be presented in a companion paper, including highlighting where HAAQI models the perception of Basic Audio Quality well, and where it does not.

The availability of public domain databases limits the tasks that can be set in music challenges. CAD1/ICASSP24 was limited to mostly pop/rock music because of this. However, hearing loss is much more prevalent in older people and our listening panel has a preference for classical music. For this reason, CAD2 will extend the demix/remix task to include classical music for small string and woodwind ensembles. This has required the synthesis of new training data for woodwind quartets [63].

V. CONCLUSION

For the first time, a challenge methodology was applied to improve music for those with a hearing loss. The tasks focused on demixing and then remixing pop/rock music to allow a rebalancing of the instruments within a recording. We provided entrants with a common set of baseline tools, databases, an evaluation metric and challenge rules. While the design of the challenge built on previous demixing challenges, the addition of listeners with different hearing characteristics added complexity to the data, software baseline and the evaluation metric. A further innovation in the ICASSP24 challenge was the addition of loudspeaker listening and specified gains being applied to the separated stems before remixing. Loudspeaker reproduction made the separation of instruments more challenging due to frequency-dependent mixing of left and right signals. The gains applied to the tracks before remixing also highlighted poorer separation. The machine learning methods used to demix the signals were nearly all refinements of current state-of-the art algorithms, either HDemucs or OpenUnmix.

The Cadenza challenge series was established to grow a community that includes hearing difference in their audio machine learning. It was pleasing to see that the number of systems entered roughly doubled between the two challenges. This has been achieved by tapping into the community of researchers already working on sound demixing. The next challenge, CAD2, includes a task on lyric intelligibility. The hope is that researchers working on speech enhancement will adapt their algorithms to lyrics in music.

ACKNOWLEDGMENT

The authors thank our partners: BBC, Google, Logitech, RNID, Sonova, Universität Oldenburg. The project has ethical approval from the University of Salford (no. 0718). Data from challenges are available under a Creative Commons Attribution license [1], [2].

REFERENCES

[1] G. Roa Dabike and T. J. Cox, "Cadenza challenge (CAD1): Databases for the first cadenza challenge - Task1," 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13285384

[2] G. Roa Dabike and T. J. Cox, "Cadenza challenge (ICASSP24): Databases for ICASSP 2024 cadenza grand challenge," 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13285307

[3] J. Blacking, *Music, Culture, and Experience: Selected Papers of John Blacking*, Chicago, IL, USA: Univ. Chicago Press, 1995.

[4] R. MacDonald, G. Kreutz, and L. Mitchell, *Music, Health, and Wellbeing*, Oxford, U.K.: Oxford Univ. Press, 2013, doi: 10.1093/acprof:oso/9780199586974.001.0001.

[5] World Health Organization, "World report on hearing," *World Health Org.*, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240020481

[6] R. Hake, M. Bürgel, N. K. Nguyen, A. Greasley, D. Müllensiefen, and K. Siedenburg, "Development of an adaptive test of musical scene analysis abilities for normal-hearing and hearing-impaired listeners," *Behav. Res. Methods*, vol. 15, no. 11, pp. 1–26, 2023, doi: 10.3758/s13428-023-02279-y.

[7] B. C. Moore, "Effects of sound-induced hearing loss and hearing aids on the perception of music," *J. Audio Eng. Soc.*, vol. 64, no. 3, pp. 112–123, 2016, doi: 10.17743/jaes.2015.0081.

[8] K. Siedenburg, S. Röttges, K. C. Wagener, and V. Hohmann, "Can you hear out the melody? Testing musical scene perception in young normal-hearing and older hearing-impaired listeners," *Trends Hear.*, vol. 24, 2020, Art. no. 2331216520945826, doi: 10.1177/2331216520945826.

[9] A. Greasley, H. Crook, and R. Fulford, "Music listening and hearing aids: Perspectives from audiologists and their patients," *Int. J. Audiol.*, vol. 59, no. 9, pp. 694–706, 2020, doi: 10.1080/14992027.2020.1762126.

[10] S. M. Madsen and B. C. Moore, "Music and hearing aids," *Trends Hear.*, vol. 18, pp. 35–47, 2014, doi: 10.1177/2331216514558271.

[11] V. Looi, K. Rutledge, and T. Prvan, "Music appreciation of adult hearing aid users and the impact of different levels of hearing loss," *Ear Hear.*, vol. 40, no. 3, pp. 529–544, 2019, doi: 10.1097/AUD.0000000000000632.

[12] J. M. Vaisberg, A. T. Martindale, P. Folkeard, and C. Benedict, "A qualitative study of the effects of hearing loss and hearing aid use on music perception in performing musicians," *J. Amer. Acad. Audiol.*, vol. 30, no. 10, pp. 856–870, 2019, doi: 10.3766/jaaa.17019.

[13] M. Uys, L. Pottas, B. Vinck, and C. Van Dijk, "The influence of non-linear frequency compression on the perception of music by adults with a moderate to severe hearing loss: Subjective impressions," *South Afr. J. Commun. Disord.*, vol. 59, no. 1, pp. 53–67, 2012, doi: 10.4102/sajcd.v59i1.22.

[14] J. Ahn et al., "The influence of non-linear frequency compression on the perception of speech and music in patients with high frequency hearing loss," *J. Audiol. Otol.*, vol. 25, no. 2, pp. 80–88, 2021, doi: 10.7874/jao.2020.00276.

[15] N. B. Croghan, K. H. Arehart, and J. M. Kates, "Music preferences with hearing aids: Effects of signal properties, compression settings, and listener characteristics," *Ear Hear.*, vol. 35, no. 5, pp. e170–e184, 2014, doi: 10.1097/AUD.0000000000000056.

[16] S. M. Madsen, M. A. Stone, M. F. McKinney, K. Fitz, and B. C. Moore, "Effects of wide dynamic-range compression on the perceived clarity of individual musical instruments," *J. Acoustical Soc. America*, vol. 137, no. 4, pp. 1867–1876, 2015, doi: 10.1121/1.4914988.

[17] L. Bramsløw, G. Naithani, A. Hafez, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm," *J. Acoustical Soc. America*, vol. 144, no. 1, pp. 172–185, 2018, doi: 10.1121/1.5045322.

[18] M. A. Akeroyd et al., "The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094918.

[19] A. J. Benjamin and K. Siedenburg, "Exploring level-and spectrum-based music mixing transforms for hearing-impaired listeners," *J. Acoustical Soc. America*, vol. 154, no. 2, pp. 1048–1061, 2023, doi: 10.1121/10.0020269.

[20] J. P. Barker et al., "Open challenges for driving hearing device processing: Lessons learnt from automatic speech recognition," in *Proc. Speech Noise Workshop (invited)*, 2020, pp. 3–4. [Online]. Available: https://2020.speech-in-noise.eu/files/SPIN2020-Programme.pdf

[21] M. Liberman and C. Wayne, "Human language technology," *AI Mag.*, vol. 41, no. 2, pp. 22–35, 2020, doi: 10.1609/aimag.v41i2.5297.

[22] G. Fabbro et al., "The sound demixing challenge 2023–Music demixing track," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 7, no. 1, pp. 63–84, 2024, doi: 10.5334/tismir.171.

[23] G. Roa Dabike et al., "The first cadenza signal processing challenge: Improving music for those with a hearing loss," in *Proc. 2nd Workshop Human-Centric Music Inf. Res.*, 2023. [Online]. Available: https://ceur-ws.org/Vol-3528/paper4.pdf

[24] G. Roa Dabike et al., "The ICASSP SP cadenza challenge: Music demixing/remixing for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops*, 2024, pp. 93–94, doi: 10.1109/ICASSPW62465.2024.10626340.

[25] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Liberec, Czech Republic, Aug. 25–28 2015, pp. 387–395, doi: 10.1007/978-3-319-22482-4_45.

[26] A. Liutkus et al., "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, Berlin, Germany: Springer, 2017, pp. 323–332, doi: 10.1007/978-3-319-53547-0_31.

[27] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Latent Variable Analysis and Signal Separation*. Berlin, Germany: Springer, 2018, pp. 293–305, doi: 10.1007/978-3-319-93764-9_28.

[28] Y. Mitsufuji et al., "Music demixing challenge 2021," *Front. Signal Process.*, vol. 1, 2022, Art. no. 808395, doi: 10.3389/frsip.2021.808395.

[29] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 2, pp. 354–365, 2015, doi: 10.1109/TASLP.2015.2507858.

[30] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613-2617, 2019, doi: 10.21437/Interspeech.2019-2680.

[31] M. A. Akeroyd et al., "The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* Rhodes Island, Greece: IEEE, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094918.

[32] G. Stevens, S. Flaxman, E. Brunskill, M. Mascarenhas, C. D. Mathers, and M. Finucane, "Global and regional hearing impairment prevalence: An analysis of 42 studies in 29 countries," *Eur. J. Public Health*, vol. 23, no. 1, pp. 146–152, 2013, doi: 10.1093/eurpub/ckr176.

[33] P. von Gablenz, E. Hoffmann, and I. Holube, "Prevalence of hearing loss in northern and southern Germany," *HNO*, vol. 65, no. Suppl 2, pp. 130–135, 2017, doi: 10.1007/s00106-016-0318-4.

[34] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," 2019, doi: 10.5281/zenodo.3338373.

[35] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," in *Proc. 24th Int. Soc. Music Inf. Retrieval Conf. ISMIR*, nov. 2023, pp. 619–626, doi: *10.5281/zenodo.10265363*.

[36] A. Défossez, "Hybrid spectrogram and waveform source separation," in Proceedings of the MDX Workshop, 2021, *doi: 10.48550/arXiv.2111.03600*.

[37] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, 2019, Art. no. 1667, doi: 10.21105/joss.01667.

[38] D. Byrne and H. Dillon, "The national acoustic Laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear Hear.*, vol. 7, no. 4, pp. 257–265, 1986, doi: 10.1097/00003446-198608000-00007.

[39] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[40] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear Hear.*, vol. 20, no. 3, pp. 182–192, 1999, doi: 10.1097/00003446-199906000-00002.

[41] M. Zhou, R. Soleimanpour, A. Mahajan, and S. Anderson, "Hearing aid delay effects on neural phase locking," *Ear Hear.*, vol. 45, no. 1, pp. 142–150, 2024.

[42] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021, doi: 10.1007/s11263-021-01453-z.

[43] F. Denk, S. M. Ernst, S. D. Ewert, and B. Kollmeier, "Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles," *Trends Hear.*, vol. 22, 2018, Art. no. 2331216518779313, doi: 10.1177/2331216518779313.

[44] G. Bernardi, T. van Waterschoot, J. Wouters, and M. Moonen, "Subjective and objective sound-quality evaluation of adaptive feedback cancellation algorithms," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 5, pp. 1010–1024, May 2018.

[45] M. Torcoli and S. Dick, "Comparing the effect of audio coding artifacts on objective quality measures and on subjective ratings," in *Audio Engineering Society Convention 144*. New York, NY, USA: Audio Engineering Society, 2018.

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

[47] D. Byrne, H. Dillon, T. Ching, R. Katsch, and G. Keidser, "NAL-NL1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures," *J. Amer. Acad. audiol.*, vol. 12, no. 01, pp. 37–51, 2001.

[48] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, 2020, Art. no. 2154, doi: 10.21105/joss.02154.

[49] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 775–785, Apr. 2013, doi: 10.1109/TASL.2012.2234114.

[50] E. M. Grais, F. Zhao, and M. D. Plumbley, "Multi-band multi-resolution fully convolutional neural networks for singing voice separation," in *Proc. 28th Eur. Signal Process. Conf.*, 2020, pp. 261–265, doi: 10.23919/Eusipco47968.2020.9287367.

[51] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *IEEE 2021 Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 51–55, doi: 10.1109/ICASSP39728.2021.9414044.

[52] M. Daly, "Remixing music for hearing aids using ensemble of fine-tuned source separators," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Workshops*, 2024, pp. 109–110, doi: 10.1109/ICASSPW62465.2024.10627557.

[53] H. Lan, T. Cheng, M. He, H. Chen, and J. Du, "The USTC system for cadenza 2024 challenge," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. Workshops*, 2024, pp. 57–58, doi: 10.1109/ICASSPW62465.2024.10627147.

[54] C. Han and S. Lee, "Optimizing music source separation in complex audio environments through progressive self-knowledge distillation," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. Workshops*, 2024, pp. 13–14, doi: 10.1109/ICASSPW62465.2024.10626965.

[55] H. Yin, M. Wang, J. Bai, D. Shi, W.-S. Gan, and J. Chen, "Sub-band and full-band interactive U-Net with dprnn for demixing cross-talk stereo music," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech, Signal Process. Workshops*, 2024, pp. 21–22, doi: 10.1109/ICASSPW62465.2024.10627597.

[56] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096956.

[57] K. Shao, K. Chen, and S. Dubnov, "Music enhancement with deep filters: A technical report for the ICASSP 2024 cadenza challenge," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. Workshops*, 2024, pp. 119–120, doi: 10.1109/ICASSPW62465.2024.10626759.

[58] J. Chen, S. Vekkot, and P. Shukla, "Music source separation based on a lightweight deep learning framework (DTTNET: Dual-path TFC-TDF UNet)," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 656–660, doi: 10.1109/ICASSP48485.2024.10448020.

[59] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: A two-stream neural network for music demixing," in *Proc. MDX Workshop*, 2021, doi: 10.48550/arXiv.2111.12203.

[60] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6567–6576, doi: 10.1109/ICCV48922.2021.00650.

[61] J. L. Drever and A. Hugill, "Aural diversity: General introduction," in *Aural Diversity*. Seattle, WA, USA: Routledge. 2022, pp. 1–12, doi: 10.4324/9781003183624-1.

[62] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. on Mach. Learn.*, vol. 202, 23–29, Jul 2023. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[63] T. J. Cox and G. Roa Dabike, "Cadenza challenge (CAD2): Databases for rebalancing classical music task (1.0.0)," 2024, available online at: https://doi.org/ 10.5281/zenodo.12664932.

GERARDO ROA-DABIKE received the B.S. degree in informatics engineering from the Universidad Diego Portales, Santiago de Chile, Chile, in 2004, and the M.Sc. and Ph.D. degrees in computer science from The University of Sheffield, Sheffield, U.K., in 2016 and 2022, respectively. He was a Postdoctoral Researcher with both the University of Salford and the University of Sheffield, Sheffield. He is currently a Research Associate with the University of Sheffield, Sheffield, working on the Cadenza Project. His recent work has been focused on machine learning methods applied to improve music for people with hearing loss.

MICHAEL A. AKEROYD received the B.A. degree in natural sciences from the University of Cambridge, Cambridge, U.K. in 1991, the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K. in 1992, and the Ph.D degree from the University of Cambridge, Cambridge, U.K. in 1995. He is Professor of Hearing Sciences with the University of Nottingham, Nottingham, U.K., and previously was a Researcher with the MRC Applied Psychology Unit, MRC Institute of Hearing Research, University of Connecticut Health Centre, the University of Sussex, and the Scottish Section of the MRC Institute of Hearing Research. His main research interests are auditory impairment and disability, hearing aids, quality of life, spatial hearing, cross-talk cancellation and noise cancellation, psychophysics, speech perception, and fMRI neuroimaging. He is a Fellow of the Acoustical Society of America.

SCOTT BANNISTER received the B.A. degree in music technology from Lancaster University, Lancaster, U.K., in 2014, the M.A. degree in applied psychology of music from University of Leeds, Leeds, U.K., in 2015, and the Ph.D. degree in music psychology from Durham University, Durham, U.K., in 2020. He was a Research Governance Support Officer with the University of Manchester, Manchester, U.K., and a Teaching Fellow in Music Psychology with the University of Leeds, Leeds, U.K. He is currently a Postdoctoral Research Fellow with the University of Leeds, working on the Cadenza Project. He has authored or coauthored research in areas of music and emotion, music and hearing loss, and neuroscience. His current research interests include music perception in hearing loss, empathy, social cognition, open research, and the emotional phenomenon of "chills". Dr. Bannister is a Member of the Society for Education, Music and Psychology Research.

JON P. BARKER is currently a Professor of Speech Processing with the University of Sheffield, Sheffield, U.K. He has been working in the area of speech and audio processing for more than 30 years. His earlier work was on modelling human speech processing with a focus on computational models of speech in noise perception. He also has interests in distant microphone speech processing and was a founder of the CHiME series of challenges and workshops for distant microphone ASR and diarisation. More recently he has been trying to bring these interests together in the area of hearing aid processing with research funded via the EPSRC Clarity https://claritychallenge.org and Cadenza projects https://cadenzachallenge.org and students funded by Sheffield's UKRI CDT in Speech and Language Technology.

TREVOR J. COX was born in Bristol, U.K. in 1967. He received the BS.c. degree in physics from the University of Birmingham, Birmingham, U.K., in 1989 and the Ph.D. degree in Acoustics from University of Salford, Salford, U.K., in 1992. He is currently a Professor of acoustical engineering with the University of Salford, where he has been a member of Faculty since 1995. His research interests include architectural acoustics, hearing devices, machine learning and psychoacoustics. He was an Engineering and Physical Sciences Research Council (EPSRC) Senior Media Fellow and has presented 25 documentaries for BBC Radio. He has written two popular science books: Sonic Wonderland (Bodley Head, 2014) and Now You're Talking (Bodley Head, 2018). Dr. Cox is an Honorary Fellow and Past President of the Institute of Acoustics (U.K.).

BRUNO FAZENDA received the B.Sc. (1st Hons.) degree in audio technology in 1999 and the Ph.D. degree in acoustics and psychoacoustics in 2004, from the University of Salford, Salford, U.K. His Ph.D. was funded by the Portuguese Fundacao para a Ciencia e Tecnologia. He was a Research Fellow with a Marie Curie Research Fellowship with the Danish Technical University, Denmark, before becoming a Lecturer in the U.K. He is a Reader (Associate Professor) in the Acoustics Research Centre with the University of Salford. His research interests in acoustics, psychoacoustics, multimodal perception in virtual realities and archaeoacoustics, focus particularly on the assessment of how an acoustic environment, technology or psychological state impact on the perception of sound. He is also a keen student on aspects of human evolution, perception, and brain function. Dr. Fazenda is a Member of the Audio Engineering Society and the Institute of Acoustics.

JENNIFER L. FIRTH received the BS.c. degree in psychology from Manchester Metropolitan University, Manchester, U.K., in 2009, the M.Sc. degree in cognitive neuroscience and neuroimaging from the University of Nottingham, Nottingham, U.K. in 2013, and she is currently working toward the Ph.D. degree in psychology with Nottingham Trent University, Nottingham, . She is also a Research Assistant in Hearing Sciences with the University of Nottingham, U.K. Her main research interests are hearing aids, quality of life, ageing, and personality.

SIMONE GRAETZER graduated in 2013 with a Ph.D. degree from the University of Melbourne, Melbourne, Australia. She has been conducting research primarily in the areas of speech acoustics and psychoacoustics. She is currently Senior Research Fellow within the Acoustics Research Centre with the University of Salford, Salford, U.K. She is a Member of the Institute of Acoustics and a full member of the Acoustical Society of America. Dr. Graetzer is a CoLead in the U.K. Research and Innovation (UKRI) Engineering and Physical Sciences Research Council funded Noise Network Plus.

ALINKA GREASLEY is currently a Professor of Music Psychology with the University of Leeds, Leeds, U.K. and a Chartered Psychologist with the British Psychological Society. Her research focuses on the impact of hearing loss on music perception and appreciation and she has been the recipient of major U.K. Research & Innovation funding to pursue this work, most notably the Hearing Aids for Music project which was the first large-scale investigation of the effects of hearing loss and the use of hearing aids on music listening. She also has research interests in musicians' hearing health, hearing conservation, and audiological practice for music rehabilitation.

REBECCA R. VOS is currently a University Research Fellow in the Acoustics Research Group, in the School of Science, Engineering & Environment, with the University of Salford, Salford, U.K. Her research interests are in understanding and tackling the problems faced by hearing impaired people singing and making music. She is involved in the Cadenza Project, also in the Acoustics Research Group, and her previous roles include Research Fellow on this project, and Research Associate in the Speech and Audio Processing group with Imperial College London, working on the EPSRC-funded ELOSPHERES Project. Her previous work has studied the Singing voice, and speech enhancement for hearing aid use in noisy situations (primarily in the areas of beamforming and array processing). She is also a Member of the Institute of Acoustics Speech and Hearing special interest group.

WILLIAM M. WHITMER is currently a Senior Investigator Scientist with Hearing Sciences – Scottish Section, the University of Nottingham, Nottingham, UK. He has devoted most of the last two-plus decades of research, both in industry and academia, towards better understanding and alleviating hearing loss. His recent work has focused on evaluating new technologies in hearing-assistance devices and understanding the perceptual relevance of technological benefits for patients.