THE ICASSP SP CADENZA CHALLENGE: MUSIC DEMIXING/REMIXING FOR HEARING AIDS

Gerardo Roa Dabike[1]   Michael A. Akeroyd[2]   Scott Bannister[3]   Jon Barker[4]
Trevor J. Cox[1]   Bruno Fazenda[1]   Jennifer Firth[2]   Simone Graetzer[1]   Alinka Greasley[3]
Rebecca R. Vos[1]   William M. Whitmer[2]

[1] University of Salford, [2]University of Nottingham, [3]University of Leeds, [4]University of Sheffield

ABSTRACT

This paper reports on the design and results of the 2024 ICASSP SP Cadenza Challenge: Music Demixing/Remixing for Hearing Aids. The Cadenza project is working to enhance the audio quality of music for those with a hearing loss. The scenario for the challenge was listening to stereo reproduction over loudspeakers via hearing aids. The task was to: decompose pop/rock music into vocal, drums, bass and other (VDBO); rebalance the different tracks with specified gains and then remixing back to stereo. End-to-end approaches were also accepted. 17 systems were submitted by 11 teams. Causal systems performed poorer than non-causal approaches. 9 systems beat the baseline. A common approach was to fine-tuning pretrained demixing models. The best approach used an ensemble of models.

Index Terms— hearing aid, hearing loss, music demixing, machine learning, challenge

1. INTRODUCTION

430 million people worldwide experience disabling hearing loss, with unaddressed hearing loss having a global cost of US $980 billion per annum[1]. Hearing loss has a number of effects, including making it harder to pick out sounds from a mixture, such as the melody line from a band. This can make music less enjoyable and risks people disengaging from listening and creating music.

Hearing aids are the main treatment for hearing loss. The default settings on hearing aids are optimized for speech, however. 68% of users report difficulties when listening to music through hearing aids [2]; the effectiveness of music programmes on hearing aids varies.

The Cadenza Project is addressing this need for better music processing through machine learning challenges. The first Cadenza Challenge (CAD1) [3] started in 2023. The subsequent 2024 ICASSP Cadenza Challenge[1] built upon the CAD1 headphone task.

2. ICASSP CHALLENGE DESCRIPTION

A person is listening to music over stereo loudspeakers. The signals to be processed are from the hearing aid microphones at each ear. Entrants were challenged to rebalance the vocal, drums, bass and other (VDBO) components by specified gains. Such a system would then allow music to be personalised, for example, amplifying the vocal component to increase lyric intelligibility.

The microphone signals were a mixture of both the right and left loudspeaker signals – see Figure 1. Head-Related Transfer Functions (HRTFs) modelled the sound propagation from the loudspeakers to
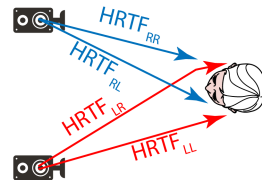
---
[1]https://cadenzachallenge.org/

Fig. 1: Left and right signals mixed using anechoic Head Related Transfer Functions (HRTFs)
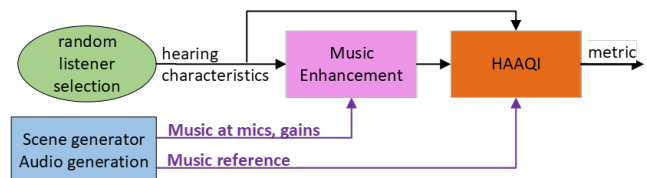

Fig. 2: Schematic of the challenge baseline

the hearing aid microphones. These were from OlHeaD-HRTF [4]. The mixing of the left-right loudspeaker signals was strongest at low frequencies where wavelengths were large compared to head size. Thus, the left-right balance of VDBO components differed in the hearing aid signals compared to the original loudspeaker feeds. Entrants who used a demix/remix approach, therefore faced additional challenge compared to CAD1 and previous demixing challenges.

Figure 2 shows a schematic of the challenge baseline. A scene generator (blue box) randomly generated scene characteristics, which included selecting the music track, choosing HRTFs characterized by the loudspeaker locations, and selecting one of the 16 subjects from the OlHeaD-HRTF dataset. Additionally, it determined the gains to be applied to each VDBO stem in the downmix. The listener audiograms were provided as metadata (green oval).

The music enhancement stage (pink box) was where an entrants' system was placed. This took the music captured by the hearing aid microphones as inputs. The output processed signals were the rebalanced stereo music.

The remixed stereo was evaluated using the Hearing-Aid Audio Quality Index (HAAQI) [5]. The reference for this intrusive metric was the rebalanced stereo using the ground truth VDBO with the appropriate HRTFs, gains and audiograms applied.

HAAQI allows for the raised hearing threshold of listeners in its calculation. The thresholds came from bilateral pure-tone audiograms at [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz. A broader bandwidth would have been preferable for music, but was not possible due to available databases and the gain rules used in HAAQI. There were 83, 51 and 53 independent pairs of measured

audiograms for the training, validation and evaluation sets.

The music for training and evaluation used the standard splits from MUSDB18-HQ [6], corresponding to 100 and 50 stereo tracks, respectively. An independent validation set was constructed by randomly selecting 50 tracks from MoisesDB [7], maintaining the same genre distribution as the evaluation split of MUSDB18-HQ. This was done because many pretrained models that use MUSDB18-HQ, incorporated the validation split as part of their training.

The two baseline systems (T01&02) were out-of-the-box pretrained audio source separation networks (with no retraining to allow for the loudspeaker scenario). T01 used the Hybrid Demucs model [8], which employs a U-Net architecture to combine both time-domain and spectrogram-based audio source separation. T02 used the Open-Unmix model [9], which just uses spectrograms.

3. SUBMISSIONS AND RESULTS

There were 17 systems from 11 teams - see Table 1. Systems that employed data augmentation or supplementation were scored before and after applying these techniques. Nearly all differences between the system scores in Table 1 are statistically significant, but some have very small effect sizes. 9 systems beat the best baseline (T01). Systems T22 and T47 scored higher using an ensemble of fine-tuned pretrained systems.

4. DISCUSSION AND CONCLUSIONS

Existing audio source separation networks like HDemucs from the Baseline T01, needed to be adapted to work with the microphone signals at the ear (as was done by T09, T09B, T11, T46). Music source separation was more difficult for this scenario, however, due to the frequency-dependent mixing of the VDBO components across the left-right channels.

The entrants' submissions comprised different techniques, ranging from ensembles of two or more audio source separation networks (T47), to the use of traditional machine learning techniques (T16). When applied with care, techniques like data augmentation and data supplementation were found to aid model generalisation (T03, T11, T31, and T42), but the gains in HAAQI were modest. Another insight from T18, was that extracting instrument-based sub- and full-band information in conjunction with beamforming can help.

In the long run, a system that is deployable on a hearing-aid needs to have a small model size, be causal and be personalised to a listeners' hearing acuity.

System T16 employed traditional machine learning technologies that required lower resources. Listener's ear embedding was explored by T09, letting the model learn to apply the amplification. T09, T09B and T16 used causal approaches. These all performed worse than all the non-casual approaches, however.

Future audio source separation challenges therefore need to encourage causal and low-latency approaches. This would enable those with hearing loss to benefit from the latest audio machine learning.

Future Cadenza challenges will also address other issues that listeners using hearing aids have with music, such as coping with large dynamic ranges without introducing distortion.

5. ACKNOWLEDGEMENTS

Table 1: System results for the evaluation set. *: causal system. A: data augmentation. S: supplementary data. B: second submission.

Entry	HAAQI	Entry	HAAQI	Entry	HAAQI
T47 [10]	**0.632**	T12	**0.573**	T31	0.530
T22	**0.631**	T46 [14]	**0.570**	T02	0.511
T03S [11]	**0.593**	T01	0.570	T09B*	0.479
T03 [11]	**0.592**	T25	0.561	T09*	0.478
T11A [12]	**0.586**	T31A	0.543	T16*	0.144
T18 [13]	**0.585**	T42	0.543		
T11 [12]	**0.580**	T42A	0.534		

6. REFERENCES

[1] World Health Organization (WHO), "Deafness and hearing loss," 2023.

[2] Alinka Greasley, Harriet Crook, and Robert Fulford, "Music listening and hearing aids: perspectives from audiologists and their patients," *Int. J. Audiol.*, 2020.

[3] Gerardo Roa Dabike, Scott Bannister, Jennifer Firth, Simone Graetzer, Rebecca R. Vos, Michael A. Akeroyd, Jon Barker, Trevor J. Cox, Bruno Fazenda, Alinka Greasley, and William M. Whitmer, "The First Cadenza Signal Processing Challenge: Improving Music for Those With a Hearing Loss," in *HCMIR*, 2023.

[4] Florian Denk, Stephan M. A. Ernst, Stephan D. Ewert, and Birger Kollmeier, "Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles," *Trends Hear.*, 2018.

[5] James M Kates and Kathryn H Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015.

[6] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," 2019, 10.5281/zenodo.3338373.

[7] Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," 2023, arXiv preprint arXiv:2307.15913.

[8] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.

[9] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-Unmix - a reference implementation for music source separation," *J. Open Source Softw.*, 2019.

[10] Matthew Daly, "Remixing music for hearing aids using ensemble of fine-tuned source separators," in *IEEE ICASSP*, 2024.

[11] Hongbo Lan, Tianyou Cheng, Maokui He, Hang Chen, and Jun Du, "The USTC System for Cadenza 2024 Challenge," in *IEEE ICASSP*, 2024.

[12] ChangHeon Han and SuHyun Lee, "Optimizing music source separation in complex audio environments through progressive self-knowledge distillation," in *IEEE ICASSP*, 2024.

[13] Han Yin, Mou Wang, Jisheng Bai, Dongyuan Shi, Woon-Seng Gan, and Jianfeng Chen, "Sub-band and full-band interactive U-NET with DPRNN for demixing cross-talk stereo music," in *IEEE ICASSP*, 2024.

[14] Keren Shao, Ke Chen, and Shlomo Dubnov, "Music Enhancement with Deep Filters: A Technical Report for The ICASSP 2024 Cadenza Challenge," in *IEEE ICASSP*, 2024.