MUSIC DEMIXING AND REBALANCING FOR HEARING LOSS

Ananya Bhardwaj **Michael Mendelson** **Ray Fairbank** **Jerry Huang**
Georgia Institute of Technology
Audio Content Analysis Fall 2023
ab22@gatech.edu, mmendelson3@gatech.edu, rfairbank3@gatech.edu, jhuang633@gatech.edu

ABSTRACT

Over 430 million people worldwide have Hearing Loss (HL), which affects their music listening experience. In this study, a signal processing and machine learning pipeline is developed to personalize music to the preferences and hearing characteristics of HL listeners. This includes utilizing machine learning source separation models (OpenUnmix, Demucs) for binaural audio demixing, VDBO stem reweighting based on listener preference, hearing aid processing (NAL-R), and evaluation with a perceptual metric of Hearing Aid Audio Quality Index (HAAQI). The derived insights span the effects of HRTFs on music demixing models as well as the impact impact of hearing aid processing and stem reweighting on HAAQI evaluation. The results should inform the development of better signal processing solutions such as demixing models for binaural audio and artifact-free stem reweighting to benefit HL listeners.

1. INTRODUCTION

1.1 Music Listening for those with Hearing Loss

According to the World Health Organization's (WHO) Report on Hearing, there are about 466 million adults worldwide with disabling hearing loss [10]. This means that they have a hearing loss of at least 40 decibels in the better hearing ear, which is equivalent to the sound of normal conversation. Additionally, the report also states that by 2050, nearly 2.5 billion people will have some degree of hearing loss. Hearing loss can negatively impact the experience of listening to and creating music. It may be difficult for those with hearing loss to hear certain frequencies or make out lyrics in music. For this reason, it is important for audio content creators, audio technology producers as well as musicians to develop solutions which do not exclude this large population, but make music and audio as inclusive and adaptable as possible to those with hearing impairments.

In this work, we will introduce an approach to adaptively adjust incoming music for hearing aid users. We use a customizable weighting of stem gains along with hearing aid processing based on individual hearing loss characteristics (via audiograms) to create an end-to-end music rebalancing system that allows the hearing impaired to adjust music for optimal listening and enjoyment.

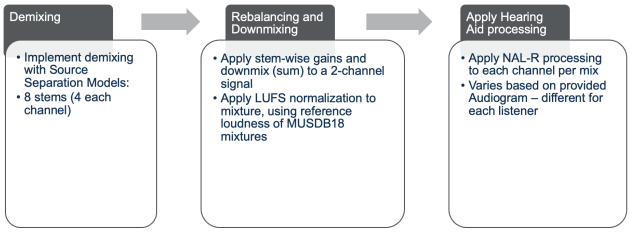

Figure 1. *Overview of approach.*

2. METHODOLOGY

2.1 Audio Processing Pipeline

In order to adaptively and customizably remix music for the hearing impaired, we employ a three step approach (Fig. 1). First, we demix our input audio into voice, drums, bass, and other (VDBO) stems. After this, we apply stem-wise gains (which are user customizable), downmix these to a stereo signal, and apply LUFS normalization to our mixture. Finally, we apply National Audio Lab's NAL-R hearing aid processing [**?**] to each stereo channel, which customizes the output based on the audiogram for each individual. In this section, we will outline our methodology for adaptively remixing audio in this way. All code to recreate our approach, as well as data and processed audio samples, are available on our github (`https://github.com/22ananya/MUSI6201`).

2.1.1 Dataset and Inputs

In order to train a music source separation model, the ICASSP 2024 Cadenza Challenge [1] outlines the use of the MUSDB18HQ [9] training dataset (100 tracks) as the primary source of music. To incorporate the effects of the listener physiology, the Head Related Transfer Functions (HRTFs, N = 96) are provided, and we utilize the provided baseline script to generate a training dataset of 800 at-ear binaural tracks (100 MUSDB tracks, convolved with 8 HRTFs each). In addition to the training set, a validation set of 967 10-second binaural music tracks are provided (based on the MUSDB18 evaluation set). Additionally, the provided dataset includes listener data (83 unique audiograms), which has the preferred audio stem gains in LUFS

and Audiograms in dB HL of a set of listeners (N = 1105 total permutations of audiograms and stem weights). A script is provided to generate scene–listener (HRTF/track-audiogram/stem weights) pairs, based on which the audio is to be enhanced and evaluated. In all of our results, we report from a subset of 100 tracks (from the Validation set) to limit computation time.

2.1.2 Demixing

To implement demixing of the 2-channel audio mixtures into 8 VDBO stems (4 stems per channel), we chose the OpenUnmix umxl model [12] and the Demucs v4 model [6]. The selection of these demixing models was based on performance, use by challenge makers as reference, and ease of implementation of pre-trained models. We used the pre-trained models considering their complexity and the required compute resources needed for training or fine-tuning (which are far above what we have available).

OpenUnmix is a spectrogram-based approach in which 3-layer bidirectional LSTMs are trained for each VDBO target [12]. Conversely, Demucs adopts a hybrid waveform/spectrogram approach, which uses a U-net convolutional architecture where the inner-most layers are replaced by a cross-domain transformer encoder [6].

2.1.3 Rebalancing and Remixing

The demixed stems are then reweighted by adding the listener specific LUFS gains. The loudness of each individual stem is calculated using PyLoudNorm [11], then the target gain is achieved by using the PyLoudNorm normalize loudness function. For remixing, the reweighted stems are simply summed together, and the summed mixture is again renormalized to the loudness of the original unweighted mixture.

2.1.4 Hearing Aid Processing

To replicate the Hearing Aid (HA) processing, we utilize the PyClarity [5] python package's provided National Acoustics Lab NAL-R hearing aid fitting formula code [**?**, 4]. It generates a FIR filter to apply frequency based linear gain, using interpolation on the frequency based HL values in the listener audiograms. The HA code is different for each ear, and operates on the left and right channel audio mixtures separately. The effect of HA processing and an example audiogram can be seen in Figure 2.

2.2 Audio Evaluation

In order to evaluate our approach, we use the Hearing-Aid Audio Quality Index (HAAQI) [7] as outlined by the ICASSP Cadenza Challenge. HAAQI is an intrusive method which compares our processed signal with a reference signal to predict music quality for individuals listening through hearing aids. The index is constructed by combining a term sensitive to noise and nonlinear distortion with a second term sensitive to changes in the long-term spectrum.

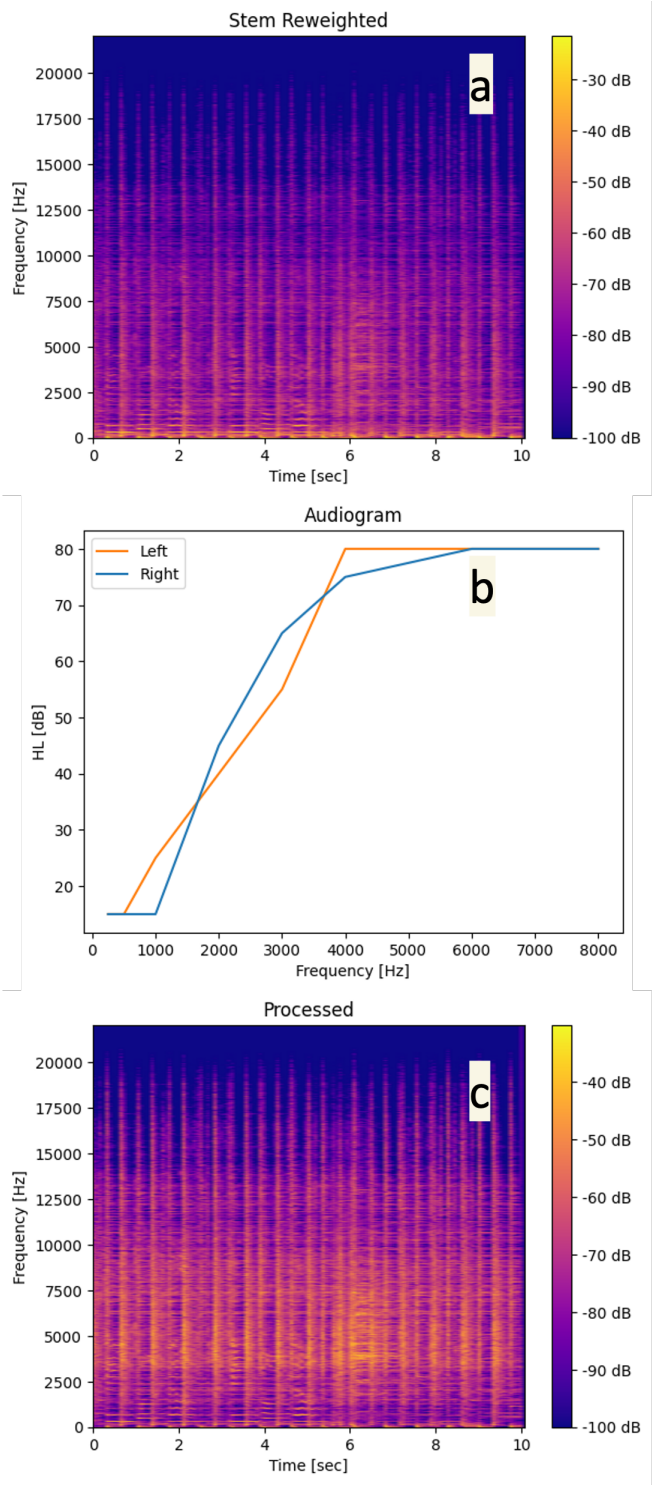

Figure 2. *Hearing Aid Processing.* The effects of Hearing Aid processing are showed in figure. a) The spectrogram of a stem reweighted signal is shown at top. b) the audiogram of a sample listener in dB HL is shown. c) The result of Hearing Aid processing is shown. The NAL-R algorithm [**?**] applies a linear frequency weighted gain based on the dB of hearing loss in the Audiogram.

HAAQI is fit to the data of [2], which is an extensive quality-rating experiment wherin music subjected to 100 different signal processing conditions representative of hearing-aid use is subjectively ranked by healthy and HL

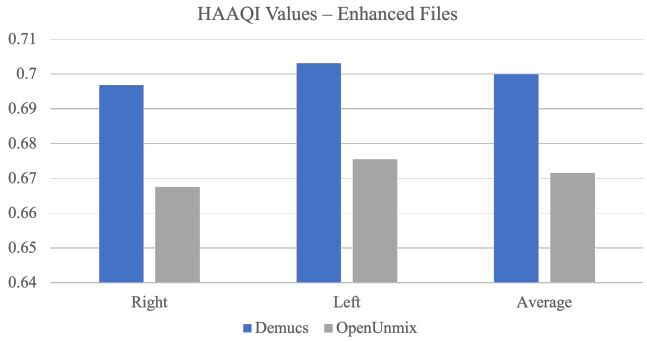

Figure 3. *HAAQI Evaluation*

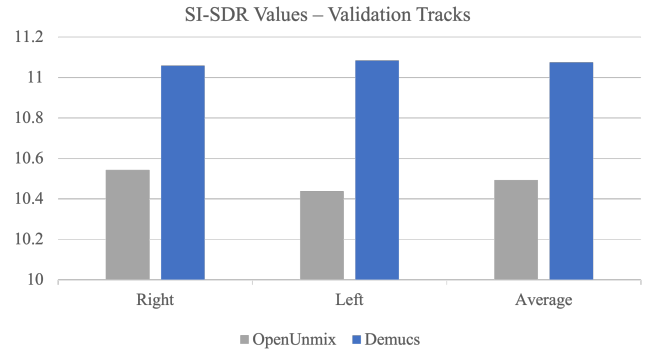

Figure 4. *SI - SDR Evaluation*

individuals. The study found that both groups gave similar rankings, which were most strongly effected by noise and nonlinear distortion. We aim to maximize HAAQI, which characterizes a signal's closeness with highly ranked signals.

3. RESULTS

As objective measures of the model performance, we used HAAQI (Fig. 3) and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [8] (Fig. 4). Due to the differences between the L and R channels (different audiograms, stem weights, signals) we evaluated both channels separately. The Demucs and OpenUnmix models could be compared since the rest of the processing chain was identical. Our results demonstrate that for both measures, Demucsv4 significantly outperforms the OpenUnmix UMXL model, and any inter-channel differences for the same model are not statistically significant. Whereas Demucsv4 has a better demonstrable performance in source-separation tasks compared to OpenUnmix [6], it doesn't necessarily suggest better performance in our evaluation, since the model applies additional signal processing after source separation, and additionally the HAAQI measure incorporates perceptual metrics for varying users with different HL. The interplay of specific demixing artefacts with the additional signal processing and perceptual measures was not known prior to this study.

4. DISCUSSION

Our signal processing pipeline of demixing $\rightarrow$ reweighting and remixing $\rightarrow$ hearing aid processing performs well in both subjective (internal listening tests the authors conducted) and objective measures (HAAQI and SI SDR evaluations of the final mixtures). Some conclusions we drew from this project were:

- Current SOTA demixing models perform perceptually well even on binaural (HRTF convolved mixtures) - as far as our subjective listening / selected objective measures go.
- In our estimation, the greatest area of potential benefit lies in development of computationally lighter, causal DSP/ML models which can operate in real-time, and maximize the perceptual audio quality specifically for HA listeners, which is a uniquely different optimization goal to that of optimizing for normal hearing individuals. HA processing can potentially amplify specific audio artifacts, and conversely, auditory perception of HL individuals can mask certain other ones. This is the primary area of opportunity.
- We relied on two Demixing models trained on very large datasets, both UMXL and Demucsv4 utilize large, private stem datasets. Additionally, both these models are non causal and cannot operate in real-time.

5. FUTURE WORK

There are multiple different thrusts of great potential improvements to extend this work. Primary among which is the training or fine-tuning of demixing models with binaural mixtures, instead of the reliance on pretrained models trained on stereo soundtracks. Additionally, newer model architectures which can incorporate HRTFs as well as binaural mixes as simultaneous inputs can be developed to perform "HRTF-aware" source separation. Secondly, since the challenge provided a predetermined baseline for evaluation, we did not investigate different approaches for each evaluation or processing step. Beginning with remixing audio stems down to a mixture, the baseline utilized simple summation of the audio signals, however, the exploration of the most suitable downmixing approach which minimizes artefacts in the summed mixture, such as frequency weighted summation may be beneficial. On hearing aid processing, the algorithm used here, the NAL-R is already been supplanted in the industry with improved non-linear fitting algorithms such as the NAL-NL1 and NL2. Utilizing the more modern HA algorithms can lead to development of a better quality implementation. In terms of evaliation metrics, different perceptual quality evaluation metrics can also be incorporated in addition to HAAQI and SI SDR that we relied upon. Going beyond the reliance of these objective metrics, exploration of actual perceived audio quality through subjective listening tests by HL listeners is the true test for any audio processing solution, using an evaluation process such as MUSHRA [3]. We suggest these further areas as aspects to improve upon a future iteration of this work.

6. REFERENCES

[1] The ICASSP 2024 Cadenza Grand Challenge | The Cadenza Project — cadenzachallenge.org. http://cadenzachallenge.org/docs/icassp_2024/intro#references. [Accessed 16-09-2023].

[2] Kathryn H Arehart, James M Kates, and Melinda C Anderson. Effects of noise, nonlinear processing, and linear filtering on perceived music quality. *International Journal of Audiology*, 50(3):177–190, 2011.

[3] ITU Radiocommunication Assembly. Itu-r bs. 1534-3:"method for the subjective assessment of intermediate quality level of audio systems,", 2015.

[4] Denis Byrne, Harvey Dillon, Teresa Ching, Richard Katsch, and Gitte Keidser. Nal-nl1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures. *Journal of the American academy of audiology*, 12(01):37–51, 2001.

[5] Gerardo Roa Dabike, Scott Bannister, Jennifer Firth, Simone Graetzer, Rebecca Vos, Michael A Akeroyd, Jon Barker, Trevor J Cox, Bruno Fazenda, Alinka Greasley, et al. The first cadenza signal processing challenge: Improving music for those with a hearing loss. *arXiv preprint arXiv:2310.05799*, 2023.

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

[7] James M Kates and Kathryn H Arehart. The hearing-aid audio quality index (haaqi). *IEEE/ACM transactions on audio, speech, and language processing*, 24(2):354–365, 2015.

[8] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[9] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017.

[10] Kaloyan Kamenov Shelly Chadha and Alarcos Cieza. The world report on hearing. *Bulletin of the World Health Organization*, 2021.

[11] Christian J Steinmetz and Joshua Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[12] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667, 2019.