Audio Mixing and Stem Proportions Adjustment

Epri Wahyu Pratiwi[1]*, Yu Tsao*[1]*, Stefano Rini*[2]

[1]Academia Sinica, Taiwan
[2]National Yang Ming Chiao Tung

epripratiwi@citi.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw, stefano.rini@nycu.edu.tw

Abstract

Hearing aid users suffer from limitations in music perception and thus prefer music with a clear audibility of lyrics. Our focus is on addressing this preference to create a more accessible and enjoyable musical experience for individuals with hearing impairments. The focus is on improving this preference through the implementation of a Demucs model for music source separation, featuring a novel proportional remixing step. This step dynamically adjusts the proportions of separated drums, bass, and vocals in the final remix based on their sound presence, utilizing an energy-based threshold approach. The algorithm covers both stem presence detection and proportional adjustment. Evaluation using Cadenza Challenge data reveals that the proposed method excels in subjective scores for multiple songs, emphasizing the positive impact of enriching vocal information on music perception. This integrated approach aims to create a more accessible and enjoyable musical experience for individuals with hearing impairments.

Index Terms: music source separation, hearing aid, music quality, audio mixing, stem proportions

1. Introduction

Audio mixing and stem proportions adjustment for hearing aid users involve techniques and considerations aimed at enhancing the auditory experience for individuals with hearing impairments. In the context of hearing aid users, audio mixing refers to the process of combining and balancing different audio elements, such as vocals, instruments, and background sounds, to optimize the overall listening experience. This is crucial for users with hearing aids, as these devices are designed to amplify specific frequencies to compensate for hearing loss.

Stem proportions adjustment, on the other hand, pertains to the individual adjustment of different audio stems or components within a piece of music. Common stems include vocals, drums, bass, and other instruments. For hearing aid users, adjusting these proportions can be particularly beneficial, as it allows for a personalized and tailored listening experience. For example, certain users may need a higher proportion of vocals to enhance speech intelligibility, while others may prefer a balance that emphasizes instrumental elements.

In music enhancement for hearing aid users, understanding the specific characteristics of the listener's hearing loss is crucial [1]. Audiograms, which depict an individual's hearing thresholds across different frequencies, play a key role in guiding the adjustment of audio elements. Additionally, considering external factors such as the listening environment, background noise, and the use of hearing aids in specific scenarios (e.g., in a car) becomes essential for creating an optimal audio mix.

Music source separation is a challenging task in audio processing, aiming to isolate individual instruments and vocals from a mixed audio track. The Demucs [2] model has demonstrated remarkable performance in this area, making it an ideal candidate for further research into music remixing. This paper combines Demucs' capabilities for source separation with an additional step for proportional remixing based on stem presence detection. In the context of enhancing audio for hearing aid users, this two-step process becomes integral. After successfully isolating the individual stems using Demucs, the proportional remixing step allows for personalized adjustments based on the user's auditory profile. For instance, the proportions of vocals, drums, bass, and other elements can be fine-tuned to accommodate specific hearing preferences, making the audio mix more tailored and user-centric. This approach not only addresses the challenges of source separation but also takes a significant stride towards optimizing the auditory experience for individuals with hearing impairments.

This paper presents how to enhance music music quality based on Task 1 and Task 2 of First Cadenza Challenge [3]. Task 1: Headphones is a challenge that centers on enhancing music source separation for individuals with hearing loss using headphones. Participants are invited to contribute by adapting a provided baseline and utilizing pretrained models to decompose stereo songs into components like vocals, drums, bass, and others (VDBO). The goal is to create a personalized remix with improved audio quality for headphone users without hearing aids. The competition comprises two stages: an enhancement stage where participants modify the baseline script, generating eight mono stems and one stereo remix for a target listener, and an evaluation stage utilizing hearing aid audio quality index (HAAQI) [4], for scoring based on comparison with reference stems. The challenge also includes a subjective evaluation with a panel of 50 listeners, and participants receive feedback on their system's performance. Overall, Task 1 aims to foster innovation in music source separation and remixing to benefit individuals with hearing loss using headphones.

Task 2: Car challenges participants to enhance music for individuals with hearing loss and hearing aids while listening in a noisy car environment. Participants are encouraged to adapt a provided baseline, modify the enhancement script, and experiment with their ideas to improve audio quality. After an objective evaluation, a subjective evaluation is conducted using a listener panel of individuals with hearing loss, and results are shared for potential use in research papers. The scenario involves a person with hearing loss wearing hearing aids in a car and listening to music played over the car stereo. The enhancement stage tasks participants with processing the music to enhance audio quality considering the presence of car noise. Access to a music dataset, listener characteristics, car speed, and

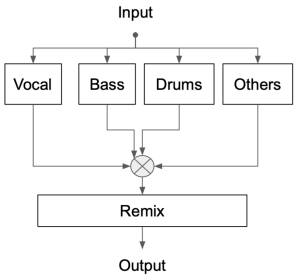

Figure 1: *Design of Experiment*

SNR at hearing aids is provided. The output of the enhancement stage is one stereo signal. The evaluation stage employs a fixed evaluation script, scoring output signals using the HAAQI method, involving the generation and application of car noise based on metadata parameters. The final output is a CSV file with HAAQI scores. The challenge emphasizes improving music listening experiences in a car for individuals with hearing loss and provides necessary datasets and information for participants.

2. Proportional Remixing and Adjustment

In this section, we introduce how to conduct the proportional remixing and adjustment.

2.1. Hearing Aid Audio Quality Index

Hearing Aid Audio Quality Index (HAAQI) is "intrusive" measurement to compare the degraded signal being evaluated to a reference signal [4]. The index is based on a model of the auditory periphery that includes the effects of hearing loss. The metrics is scalar between 0 and 1, where a higher score of HAAQI represents better audio quality. This signal is then passed to a simple hearing aid composed of a NAL-R amplification [5].

For Task 1, involving headphone listening, the reference signals for HAAQI consist of the original left and right channels of the music tracks. In Task 2, pertaining to car listening scenarios, the references are the left and right signals captured at the ear canal of a listener experiencing the music through a stereo setup with two loudspeakers, within an anechoic room.

2.2. Demucs Music Source Separation

The Demucs model [2] is a deep learning-based architecture, built upon a U-Net convolutional structure. It is designed to separate drums, bass, and vocals from a given music track. The input audio from Cadenza Challenge Task 2 is processed through the pretrain Demucs model to obtain separated drums, bass, and vocals stems, as well as other accompaniment stems.

2.3. Stem Presence Detection

The next step is detecting whether each stem contains sound or is silent. This is achieved using a simple energy-based approach. The steps determine if a given waveform (stem) contains sound by checking if its maximum amplitude exceeds a specified threshold. If the maximum amplitude is greater than the threshold, the stem is considered to contain sound.

2.4. Proportional Adjustment

These steps explain the optimization of music remix proportions to minimize the distance to the higher score. Initially, we sampled several audios that had already been separated into vocals, bass, drums, and others. We then mixed them based on specified proportions and calculated HAAQI loss. However, due to the computational consuming nature of calculating HAAQI, we alternatively used short-time objective intelligibility (STOI) [6], which ranges between 0 and 1, where a higher STOI score represents better intelligibility of lyrics. The reason we used STOI is to mimic that if a song has good quality, it will have lower loss. Optimization was achieved through methods like Powell [7], CG [8], BFGS [8], ensuring remix proportions sum to 1. Results demonstrate three sets of optimal proportions ([0.25 0.25 0.25 0.25]) with consistently high scores (0.889), for vocals, bass, drums, and others, respectively. This indicates the effectiveness of the proposed approach for achieving quality audio remixes. Initially, we assumed that the optimization would have different proportions. However, because we wanted to amplify the vocals to improve the audibility of the lyrics in a song, we decided to set our own proportions. This strategic decision aligns with findings from prior research [9], indicating that existing music pre-processing methods aim to enhance music signals for hearing-impaired listeners, especially in the case of cochlear implant users, by either emphasizing preferred voices or reducing the spectral complexity of the signals

After determining the presence of sound in each stem, the algorithm calculates new proportions for the remix. The function `adjust_proportions(silent_stems)` takes a list of binary values indicating whether each stem is silent (1) or contains sound (0). Depending on the presence of sound, the proportions of the vocal, bass, drum, and other stems are adjusted as follows. If all stems are silent, the original mix (equal proportions) is used. If at least one stem contains sound, the following proportions are applied: Vocals (60% if not silent, 0% if silent), Bass (15% if not silent, 0% if silent), Drums (10% if not silent, 0% if silent), and Other (15% if not silent, 0% if silent).

2.5. Remixing Process

The proposed remixing process is demonstrated using Python and relevant audio processing libraries. The audio paths are loaded from a list file, and for each audio path, the presence of sound in each stem is determined. The proportions are adjusted using the function described in Section 2.4. The stereo sound of each stem is modified according to the calculated proportions, and the modified stereo sounds are combined to create the final mix. The mix is normalized to ensure it doesn't exceed the maximum amplitude. The resulting remix is saved as a .flac file in the output subfolder. The detailed experiment is available at: `https://github.com/epriwahyu/cadenza_challenge`.

3. Experiments

The music for evaluation data for Task 1 and Task 2 of Cadenza Challenge were used in this experiment. In Task 1, there are 53 songs and 49 listeners. In total, 2597 combination of song and listeners for the evaluation. In Task 2, there are 70 unique scenes, 70 unique songs, 7 unique genres, and 53 unique listeners. In total, there are 3710 songs simulated, generated through combinations of scenes, songs, genres, and listeners.

In our submission, we simply implemented the proportion adjustment as explained on Section 2.5. In the evaluation

Song Title	Average (%)	
	Baseline	Ours
Al James - Schoolboy Fascination	39.1	**46.1***
Angels In Amplifiers - I'm Alright	**47.8**	43.6
BKS - Too Much	**50.1**	11.8
Carlos Gonzalez - A Place For Us	**48**	43.9
Enda Reilly - Cur An Long Ag Seol	**55.9**	32.9
Forkupines – Semantics	**34.8**	17.6
Girls Under Glass - We Feel Alright	**26.2**	23.7
Hollow Ground - Ill Fate	**28.1**	20.7
James Elder & Mark M Thompson - The English Actor	21.3	**28.7***
Little Chicago's Finest - My Own	**52.6**	34.2
M.E.R.C. Music - Knockout	**43**	26.3
Moosmusic - Big Dummy Shake	25.4	**33.9***
Motor Tapes - Shore	**38.1**	37.1
Mu - Too Bright	**34.1**	29.9
Nerve 9 - Pray For The Rain	32.5	**38.0***
PR - Happy Daze	**54.6**	43.1
PR - Oh No	**61.4**	44.3
Raft Monk – Tiring	13.9	**25.5***
Side Effects Project - Sing With Me	**42**	39.9
Speak Softly - Broken Man	39.9	**44.1***
The Doppler Shift – Atrophy	**49.9**	46.1
The Mountaineering Club - Mallory	**62.6**	46.3
The Sunshine Garcia Band - For I Am The Moon	39.6	**50.9***
Tom McKenzie - Directions	**60.5**	45.3
Triviul feat. The Fiend – Widow	33.7	**38.4***
	41.404	35.7

Figure 2: *Result listening Task 1. (*) means our remixed songs are preferred*

of Task 1, as illustrated in Table 1, the baseline achieved an HAAQI score of 0.2550 and a listening score of 41.40, while our proposed method demonstrated improvement with a lower HAAQI score of 0.1951 and a reduced listening score of 35.70. For Task 2, the evaluation results are summarized in Table 2, showcasing the objective HAAQI scores. The baseline recorded an HAAQI score of 0.1256, whereas our approach achieved a lower performance with an HAAQI score of 0.1187. In contrast, there are 8 out of 25 songs ara having better music quality by using our proportion, as can be seen on Figure 2. It shows that enhancing more information of vocal in a music is helping to improve the music perception.

	HAAQI	**Listening Score**
Baseline	0.2550	41.40
Ours	0.1951	35.70

Table 1: *Result of objective and subjective scores of Task 1*

	HAAQI
Baseline	0.1256
Ours	0.1187

Table 2: *Result of objective scores of Task 2*

4. Conclusions

This paper introduces a Demucs model implementation for music source separation, featuring a proportional remixing step that dynamically adjusts stem proportions based on sound presence. The proposed algorithm includes stem presence detection and proportional adjustment. Evaluation using Cadenza Challenge data for Task 1 and Task 2 reveals that, despite the objective score not reaching the baseline, the proposed method achieves better subjective scores for several songs. This underscores the positive impact of enhancing vocal information on music perception.

5. References

[1] A. Greasley, H. Crook, and R. Fulford, "Music listening and hearing aids: perspectives from audiologists and their patients," *International Journal of Audiology*, vol. 59, no. 9, pp. 694–706, 2020.

[2] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," 2022.

[3] G. R. Dabike, S. Bannister, J. Firth, S. Graetzer, R. Vos, M. A. Akeroyd, J. Barker, T. J. Cox, B. Fazenda, A. Greasley *et al.*, "The first cadenza signal processing challenge: Improving music for those with a hearing loss," *arXiv preprint arXiv:2310.05799*, 2023.

[4] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (haaqi)," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 2, pp. 354–365, 2015.

[5] D. Byrne and H. Dillon, "The national acoustic laboratories'(nal) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257–265, 1986.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[7] M. J. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *The computer journal*, vol. 7, no. 2, pp. 155–162, 1964.

[8] J. Nocedal and S. J. Wright, "Conjugate gradient methods," *Numerical optimization*, pp. 101–134, 2006.

[9] B. Lentz, A. Nagathil, J. Gauer, and R. Martin, "Harmonic/percussive sound separation and spectral complexity reduction of music signals for cochlear implant listeners," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8713–8717.