THE USTC SYSTEM FOR CADENZA 2024 CHALLENGE

Hongbo Lan[†], *Tianyou Cheng*[†], *Maokui He, Hang Chen, Jun Du*[‡],

University of Science and Technology of China, Hefei, P.R.China

ABSTRACT

This paper reports our submission to the ICASSP 2024 Cadenza Challenge, focusing on the non-causal system. The challenge aims to develop a signal processing system that enables personalized rebalancing of music to improve the listening experience for individuals with hearing loss when they listen to music via their hearing aids. The system is based on the Hybrid Demucs model. We fine-tuned the baseline model on the given dataset with a multi-target strategy and added a mixture of information in the downmixing stage. We combined different models and achieved a score of 0.6878 in the Hearing Aid Audio Quality Index (HAAQI) metric on the validation set and 0.5929 on the test set.

Index Terms— Music separation, Hearing aids, Rebalance, Music quality

1. INTRODUCTION

The World Health Organization reports that over 1.5 billion people worldwide suffer from hearing loss. Hearing loss can cause difficulties in listening to music, making it harder to discern lyrics and causing music to sound duller as high-frequency disappear. It is important to consider diverse hearing conditions and ensure that individuals with hearing loss can benefit from the latest signal processing advancements.

The Cadenza 2024 Challenge [1], held within the framework of ICASSP 2024, aims to improve the sound quality of music for individuals experiencing hearing impairments. The training dataset for the challenge consists of 400 songs simulated using Head-Related Transfer Functions (HRTF) based on MUSDB18-HQ [2]. Additionally, audiograms for 83 individuals with hearing loss are provided to allow participants to personalize it and improve its assistance for those with hearing loss.

As Hybrid Demucs [3] represents the current state-of-the-art open-source music separation model on the MUSDB18-HQ dataset, our system is built on the foundation provided by Hybrid Demucs. However, it is essential to note that a distinctive aspect of this challenge is the utilization of data that does not solely comprise mixed music but also accounts for leakage influences. Therefore, our initial approach was to fine-tune the Hybrid Demucs model using simulated training data. This adaptation enables the model to effectively deal with music separation tasks, including leakage. Subsequently, following the procedural framework of this challenge, we used the outputs from the enhancement stage as optimization objectives, exploring various combinations. During the enhancement stage, the outputs of different models are fused while incorporating a certain proportion of the original mixture information to aid in separation. We also investigated the impact of simulated training set sizes on model training and submitted results for the official and supplementation datasets with 800 songs.

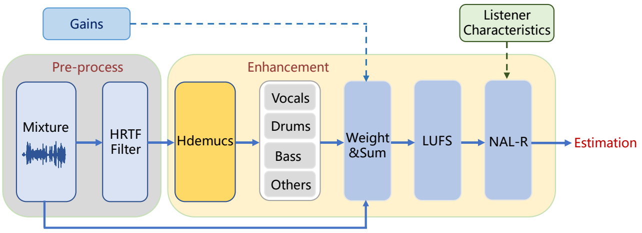

Fig. 1. The USTC system architecture for Cadenza Challenge

2. METHOD

2.1. Data supplementation and selection

In the provided dataset of 400 training songs, we randomly selected subsets of 100, 200, and 300 songs to fine-tune our system. Furthermore, we used official data supplementation tools to simulate an additional 400 songs, aiming to assess the impact of data supplementation on the separation performance of the system. As Table 1 shows, fine-tuning the model with 200 songs produces better results than 400 songs, requiring less training time. Therefore, our primary focus was on fine-tuning the model using the dataset consisting of 200 songs. In addition, fine-tuning the model with data supplementation of 800 songs can bring some improvement.

Table 1. Results of data supplementation using a different number of scenes on the validation set. “N” means the number of scenes used to fine-tune our system.

N	100	200	300	400	800
HAAQI	0.6766	0.6792	0.6781	0.6785	0.6796

2.2. System architecture

The separation model utilized in this challenge is based on the Hybrid Demucs architecture which consists of a hybrid U-Net model in the time and frequency domains. The temporal branch processes mixed music input like the standard Demucs. The spectral branch operates on the spectrogram derived from a Short-Time Fourier Transform (STFT) performed over 4096-time steps, using a hop length of 1024. The output of the spectral branch is then inverted by the Inverse Short-Time Fourier Transform (ISTFT) and combined with the output of the temporal branch. This combined result is the final prediction from the model.

This time-frequency domain hybrid model can split mixtures into four stems: vocals, drums, bass, and others (VDBO). Initially, we fine-tuned the baseline model using the simulated training set. We applied the fine-tuned separation model to the mixture of music during the enhancement stage. We applied specific gains to each

[†] equal contribution. [‡] corresponding author.

Table 2. Average HAAQI scores of different models on validation and test datasets

Model		Baseline-Hybrid Demucs	Baseline-OpenUnmix	Official-USTC	Supplementation-USTC
HAAQI	Validation	0.6677	0.5963	0.6870	0.6878
	Test	0.5697	0.5113	0.5923	0.5929

stem and completed the downmixing process. The downmixed signal is normalized based on the input mixture signal to limit the signal amplitude range. Furthermore, the National Acoustic Laboratories-Revised (NAL-R) is applied based on individualized hearing thresholds, catering to different listeners' hearing impairments. Lastly, the processed signals are evaluated against the processed reference signals to calculate the final Hearing Aid Audio Quality Index (HAAQI) metric [4].

2.3. Fine-tune models with a multi-target strategy

The system architecture consists of multiple stages, including downmixing, normalization and personalized operations for specific listeners, each producing distinct outputs. Therefore, we considered using the outputs from each enhancement stage as training objectives for the separation model. We explored both individualized training targets and their amalgamation for joint optimization. The average absolute error loss function was used between the source and estimated waveforms during training. Specifically, the average absolute error loss of each VDBO was calculated separately. Considering that the four stems were equally important, we used an average weighting method to obtain the final loss function value. As a result, we obtained models that focus on different targets, laying the groundwork for subsequent model fusion.

2.4. Post-processing for demixing stage

Expanding on Section 2.3, we have obtained models emphasizing different objectives. During the enhancement stage, we employed a subset of these models for decoding. Subsequently, these decoded results were combined with varying weights. Furthermore, the separated music may experience information loss, considering the separation model's performance limitations. To mitigate this, we augmented the separated results with a certain proportion of the original mixture. After experiments, we found that the model achieved the best result when we mixed the mixture information with a weight of 0.15.

3. EXPERIMENT

We used the Hybrid Demucs model pre-trained on the MUSDB18-HQ training dataset, with an approximate size of 320 MB. Our models were fine-tuned on 3 NVIDIA GTX-3090 GPUs with 24GB of RAM, using a batch size of 24 and an 8-second segment. All models were optimized using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an initial learning rate of 6.0×10^{-4}.

Table 2 presents the HAAQI scores of different models on the validation and test sets. The baseline contains two models, Hybrid Demucs and OpenUnmix [5], with scores of 0.6677 and 0.5933, respectively, on the validation set. The Official-USTC refers to the model trained on the dataset generated using 400 official scenes. The supplementation-USTC refers to the model trained on the additional music scenes generated by MUSDB18-HQ using official scripts.

We also observed a correlation between the HAAQI scores and the audiograms of the listeners. Those with more significant low-frequency losses tended to have lower HAAQI scores. For example, the listener L5000 in the validation dataset, whose left ear loss in the 500 - 2000 Hz range is approximately 40 dB, whereas the right ear loss in the same frequency range is over 70 dB. The HAAQI score for the listener's left ear L5000 can achieve 0.87, whereas the HAAQI values for the right ear drop to 0.22 for the same music.

4. CONCLUSION

By fine-tuning the Hybrid Demucs model on a simulated dataset enriched with HRTF, employing multi-target training, fusing different models, and incorporating mixture information in post-processing, our system achieved a HAAQI score of 0.6878 on the validation set and 0.5929 on the test set. Significant variations in the HAAQI metric for music separation were observed among hearing-impaired individuals. The model has limitations in effectively handling cases of severe hearing loss, indicating an area for future improvement and research focus.

5. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No.62171427.

6. REFERENCES

[1] Gerardo Roa Dabike, Michael A Akeroyd, Scott Bannister, Jon Barker, Trevor J Cox, Bruno Fazenda, Jennifer Firth, Simone Graetzer, Alinka Greasley, Rebecca R. Vos, and William M. Whitmer, "OVERVIEW AND RESULTS OF THE ICASSP SP CADENZA CHALLENGE: MUSIC DEMIXING/REMIXING FOR HEARING AIDS," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[2] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-hq - an uncompressed version of musdb18," Dec 2019, doi:10.5281/zenodo.3338373.

[3] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[4] James M Kates and Kathryn 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] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, pp. 1667, 2019.