Fully open-sourced music source separation and speech quality enhancement systems

Haohe Liu[1]
Doctoral Research Student
[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey
Email: hl01486@surrey.ac.uk; ORCID: 0000-0003-1036-7888

Abstract

This poster introduces a music source separation (MSS) system and a speech quality enhancement (SQE) system. Both systems are fully open-sourced.
(1) Music source separation aims to separate different sound sources (e.g., drums, bass) from a mixture audio file. MSS has several applications such as karaoke and music remixing. Our study proposed a new model to predict complex-valued ideal ratio masks with deep ResUNet architecture and channel-wise subband features. In ISMIR 2021 music demixing (MDX) challenge, our system ByteMSS ranked 2nd in the vocal track and 5th on average score.
(2)Speech quality enhancement is a crucial topic in improving online communication or recorded speech quality. Our study proposed a new model to predict complex-valued ideal ratio masks with deep ResUNet architecture and channel-wise subband features. VoiceFixer achieves state-of-the-art results on High-Fidelity speech restoration, dereverberation, de-clipping, enhancement, and equalization in one pass. Subjective evaluations show that our system has clear advantages over baselines and achieves good restoration quality on real-world test cases.

Introduction

Music source separation (MSS)
Music source separation (MSS) is a task to separate audio mixtures into individual sources such as vocals, drums, accompaniment, etc. MSS is an important topic for music information retrieval (MIR) since it can be used for several downstream MIR tasks including melody extraction, pitch estimation, , music transcription, music remixing, and so on. MSS also has several direct applications such as Karaoke and music remixing.

In this study, we propose a new MSS model, channel-wise subband phase-aware ResUNet (CWS-PResUNet), to decompose signals into subbands and estimate an unbound complex ideal ratio mask (cIRM) for each source. By combining CWS-PResUNet and Demucs, our ByteMSS system ranks the 2nd on vocals score and 5th on average score in the 2021 ISMIR Music Demixing (MDX) Challenge limited training data track (leaderboard A).

Speech Quality Enhancement (SQE)
Human speech often suffers from distortions such as background noise, room reverberations, or clipping from low-quality devices. Those distortions degrade the perceptual quality of human listeners. Speech restoration is a task to restore degraded speech to high-quality speech, which is useful in a wide range of applications such as online meeting and hearing aids.

In this work we introduce VoiceFixer, a unified framework for high-fidelity speech quality enhancement. VoiceFixer restores speech from multiple distortions~(e.g., noise, reverberation, and clipping) and can expand degraded speech~(e.g., noisy speech) with a low bandwidth to 44.1 kHz full-bandwidth high-fidelity speech. We design VoiceFixer based on (1) an analysis stage that predicts intermediate-level features from the degraded speech, and (2) a synthesis stage that generates waveform using a neural vocoder. Both objective and subjective evaluations show that VoiceFixer is effective on severely degraded speech, such as real-world historical speech recordings.

Methodology of MSS

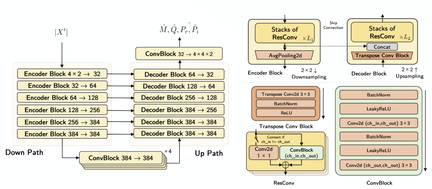


Figure 7: Proposed MSS model (CWS-PResUNet).

As is shown in Figure 7, the model is a symmetric architecture containing a down-sampling and an up-sampling path with skip-connections between the same level. It accepts magnitude spectrogram as input and estimates mask estimation, phase variations, and direct magnitude prediction. The complex spectrogram can be reconstructed with the equation on the left.

$$\hat{S}' = \mathbf{relu}(|X'| \odot \mathbf{sigmoid}(\hat{M}) + \hat{Q}) \exp^{j(\angle X' + \angle \hat{\theta})}$$

Open Research Practices

Figure 1: We open sourced our code to facilitate reproducibility. [1, 2, 3, 4]

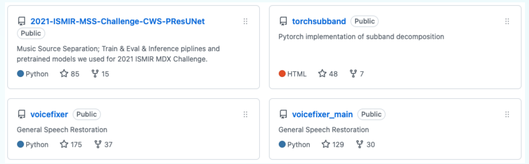


Figure 2: The PyPI packages we built for the python users. [14, 15, 16]

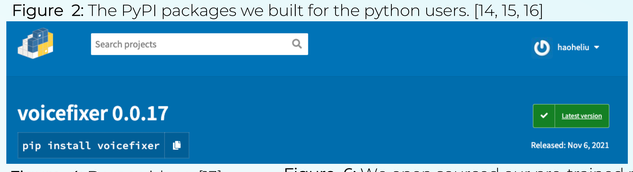


Figure 3: Utilizing google Colab for the remote reproductions of our works.

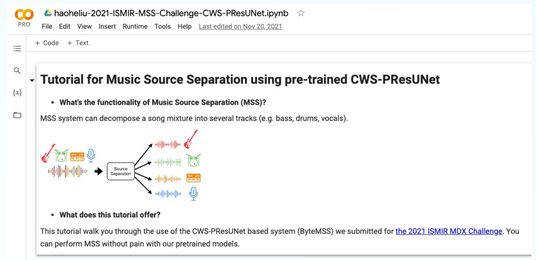


Figure 4: Demo videos. [13]

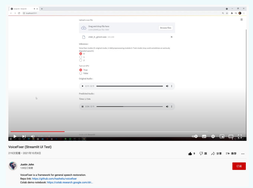
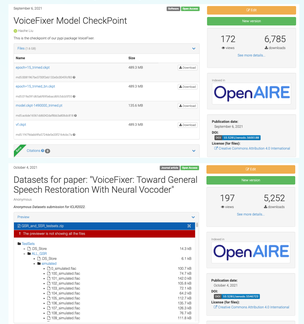


Figure 5: Web demo pages. [10, 11]

Figure 6: We open sourced our pre-trained models and datasets. [5, 6, 7]



Methodology of SQE

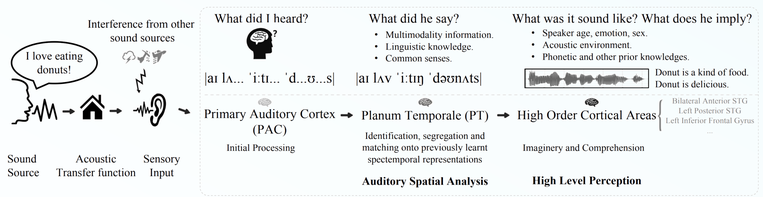


Figure 8: The neural and cognitive model of how human process distorted speech.

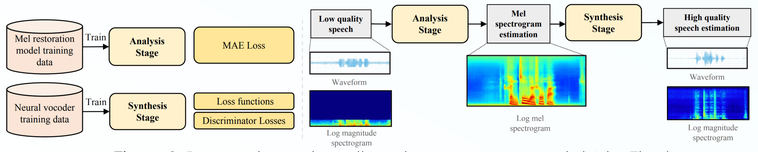


Figure 9: Proposed speech quality enhancement approach (VoiceFixer).

In speech restoration, our proposed VoiceFixer breaks the conventional one-stage system into a two-stage system:

$$f : \boldsymbol{x} \mapsto \boldsymbol{z},$$
$$g : \boldsymbol{z} \mapsto \hat{\boldsymbol{s}}.$$

The first equation denotes the analysis stage of VoiceFixer where a distorted speech **x** is mapped into a representation **z**. The second equation denotes the synthesis stage of VoiceFixer, which synthesize **z** to the restored speech. Through the two-stage processing, VoiceFixer mimics the human perception of speech described in Figure 8.

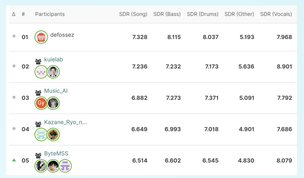
Compared to the conventional speech restoration methods that operate on spectrogram or waveform, VoiceFixer uses the low dimensional mel spectrogram as **z**, which alleviates the difficulties of restoring multiple distortions simultaneously.

Results

Our MSS system (ByteMSS) **outperforms SOTA methods** by a large margin and **ranks the 2nd in vocal separation** performance in 2021 ISMIR MDX challenge.

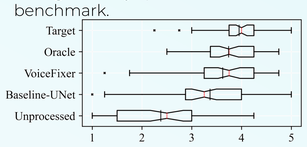
Table 1: Evaluation result on MSS benchmark MUSDB18HQ

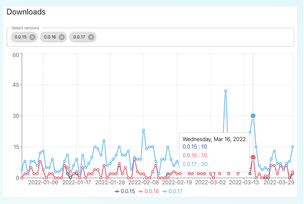
Models	Vocals	Drums	Bass	Other	Average
X-UMX	6.61	6.47	5.43	4.64	5.79
D3Net	7.24	7.01	5.25	4.53	6.01
Demucs	6.89	**6.57**	**6.53**	5.14	6.28
CWS-PResUNet	**8.92**	6.38	5.93	**5.84**	6.77
ByteMSS	8.92	6.57	6.53	5.84	**6.97**



VoiceFixer achieves **SOTA performance** on multiple speech quality enhancement benchmarks. Our open-sourced tools receive a lot of attentions from the speech researcher. **The voicefixer package on PyPI has been downloaded for more than 13k times in the last six months.**

Figure 10: Subjective evaluation result on high-fidelity speech restoration benchmark.





Relevant links

[1] https://github.com/haoheliu/voicefixer_main
[2] https://github.com/haoheliu/voicefixer
[3] https://github.com/haoheliu/2021-ISMIR-MSS-Challenge-CWS-PResUNet
[4] https://github.com/haoheliu/torchsubband
[5] https://zenodo.org/record/5546723#.YkRLiprMKAk
[6] https://zenodo.org/record/5600188#.YkRLpJrMKAk
[7] https://zenodo.org/record/5175846
[8] https://colab.research.google.com/drive/1E2yJLWN8MH6GJUw15cj490E2BaeAcGie
[9] https://colab.research.google.com/drive/1HYYUepIsI2aXsdET6P_AmNVXuWP1MCMf
[10] https://haoheliu.github.io/nvsr
[11] https://haoheliu.github.io/voicefixer
[12] https://huggingface.co/spaces/akhaliq/VoiceFixer/tree/main
[13] https://www.youtube.com/watch?v=d_j8UKTZ7J8
[14] https://pypi.org/project/voicefixer/0.0.17/
[15] https://pypi.org/project/torchsubband/
[16] https://pypi.org/project/ssr-eval/