SDX
Sound Demixing Workshop 2023

One-page Report for Tencent AI Lab's CDX 2023 System

Kai Li[1,2], **Yi Luo**[1], **Jianwei Yu**[1], **and Rongzhi Gu**[1]

1 Tencent AI Lab, Shenzhen, China **2** Tsinghua University, Beijing, China

Published: 04 November 2023

Abstract

Tencent AI Lab introduces its innovative system for the CDX 2023 challenge, emphasizing audio source separation. The research leveraged both public and Tencent's internal datasets, totaling hundreds of hours of audio data. A pivotal preprocessing step involved removing human vocal components from musical and effect data, which substantially enhanced system performance. To augment the training dataset's diversity, on-the-fly data mixing was employed, sampling and adjusting various audio signals. The Band-Split RNN (BSRNN) architecture was the backbone of the system, addressing separation tasks. Interestingly, a model tailored for music separation demonstrated superior performance in extracting dialogue compared to a speech-centric model. The system's primary strength was evident in its ability to separate the audio track, earning it the top position in the CDX challenge. This paper delves into the encountered challenges, particularly in dialogue track separation, and postulates the reasons behind the observed outcomes.

Dataset

We used the public divide and remaster (DnR) (Petermann et al., 2021) dataset, the public deep noise suppression (DNS) dataset (Dubey et al., 2022), the public MUSDB18-HQ dataset (Rafii et al., 2019), and some extra internal data for system training. The extra internal speech data include 150 hours of data used for text-to-speech task, the extra internal sound effect data include 10 hours of cinematic sound effect data, and the extra internal music data include 100 hours of cinematic background music data.

One important step in our data preprocessing pipeline was that we found that the effect and music signals in both the DnR dataset and our internal dataset may contain human voice. We thus used a music source separation (MSS) model to preprocess all the effect and music signals to subtract the extracted "speech" or "vocal" signals from them. We found that doing this significantly improved the systems' performance compared to directly using the original signals for training.

Methods

On-the-fly Data Mixing

We performed on-the-fly data mixing during training to increase the variety of the training data mixtures. For each mixture utterance, we randomly sampled 0-1 speech or vocal signal (we also treated vocal signal as a form of dialog signal in our setting), 0-2 music signals and 0-3 effect signals and rescaled each of them by a random energy of [-10, 10] dB. We truncated the signals to 3-second long and then added them up to form the mixture.

The sum of individual music and effect signals were set as the training targets for the two tracks, respectively.

Model Design

Our system consists of three independent models for the dialog, effect and music tracks, respectively. All models share a same architecture, which is the band-split RNN (BSRNN) architecture we proposed for the task of music source separation (MSS) (Luo & Yu, 2023). For dialog track, we directly use a BSRNN model trained for music source separation task instead of CDX task, as we eventually found that using an MSS model trained on music-only data that extracts the vocal track from the accompaniment can lead to better SDR score on the hidden test set than a speech-extraction model trained on speech data (please see the discussion section for more on this observation). For the effect track and the music track, we used two separate BSRNN models trained on the aforementioned dataset, while we used the MSS model to first subtract the separated dialog signal from the mixture to create a pseudo music-effect-only mixture, and then trained the two models on this mixture to perform a slightly simpler separation task. We found that this could lead to better performance than training the two models on mixtures containing dialog data, and also better than training on mixtures without speech or vocal signal.

We used the standard BSRNN architecture and we do not include detailed description here for the sake of simplicity. The band-split scheme we used for all models was identical to the one we proposed in the original literature. The number of sequence and band modeling modules in the effect and music models were 8 and 12, respectively, and the feature dimension N were set to 64 and 128, respectively.

Training Configurations

All models were trained with Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.001. We used 8 GPUs for each model with a per-GPU batch size of 2. Each training epoch contained 10k iterations, and the learning rate was decayed by 0.98 for every two epochs. We did not apply early stopping as the evaluation was done on the hidden test set, and we submitted the latest model to the grading system every day to find the best model.

Results and Discussions

Our system achieved #1 on Leaderboard B and Final Leaderboard B in the CDX challenge. Comparing with other top-ranking systems, our system performed significantly better on music track and par or slightly worse on the two other tracks, and the overall improvement mainly came from the gain from the music track. One most interesting observation we had was about the dialog track - we initially tried to treat the “dialog separation” task as a “speech enhancement” task which aims at removing any non-speech components out of the mixture, and we trained systems based on both our speech enhancement system which ranked 3rd in the 5th DNS challenge (Yu et al., 2023, 2022) and our MSS system (Luo & Yu, 2023) with the extra cinematic data. We perceptually evaluated the systems’ performance on internal movie data and found the quality of their outputs satisfying. However, all model weights trained with this fashion cannot achieve 13 dB SDR on the hidden test set, no matter how we adjusted the training pipeline or the model design. Later we tried to directly submit the original MSS system only trained on music-only data (MUSDB18-HQ and some other internal music dataset), and the performance of the dialog track on Leaderboard B suddenly reached 15 dB SDR. As we empirically found that the MSS system cannot fully eliminate nonspeech sounds when directly applied to speech data and the sound effect signal often leak to the dialog track, we suspect that the dialog track in the hidden test set may contain certain environmental sounds or noise,

possibly because that the signals were real-recorded on the film sets instead of in the studios. This assumption makes the dialog extracting task harder and less intuitive as the system is required to leave certain nonspeech sounds into the dialog track while being able to distinguish them from cinematic sound effects or music signals. This was also the reason we used the MSS system to preprocess the DnR and internal dataset to subtract speech signals instead of using a strong speech enhancement model to do so.

Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S. E., Thakker, M., Yoshioka, T., Gamper, H., & others. (2022). Icassp 2022 deep noise suppression challenge. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9271–9275.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*.

Luo, Y., & Yu, J. (2023). Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1893–1901.

Petermann, D., Wichern, G., Wang, Z.-Q., & Le Roux, J. (2021). The cocktail fork problem: Three-stem audio separation for real-world soundtracks. *arXiv Preprint arXiv:2110.09958*.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019). *MUSDB18-HQ - an uncompressed version of MUSDB18*. https://doi.org/10.5281/zenodo.3338373

Yu, J., Chen, H., Luo, Y., Gu, R., Li, W., & Weng, C. (2023). TSpeech-AI system description to the 5th deep noise suppression (DNS) challenge. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2.

Yu, J., Luo, Y., Chen, H., Gu, R., & Weng, C. (2022). High fidelity speech enhancement with band-split RNN. *arXiv Preprint arXiv:2212.00406*.