Separation of multiple singing voices into each voice is a rarely studied area in music source separation research. The absence of a benchmark dataset has hindered its progress. In this paper, we present an evaluation dataset and provide baseline studies for multiple singing voices separation. First, we introduce MedleyVox, an evaluation dataset for multiple singing voices separation. We specify the problem definition in this dataset by categorizing it into i) unison, ii) duet, iii) main vs. rest, and iv) N-singing separation. Second, to overcome the absence of existing multi-singing datasets for a training purpose, we present a strategy for construction of multiple singing mixtures using various single-singing datasets. Third, we propose the improved super-resolution network (iSRNet), which greatly enhances initial estimates of separation networks. Jointly trained with the Conv-TasNet and the multi-singing mixture construction strategy, the proposed iSRNet achieved comparable performance to ideal time-frequency masks on duet and unison subsets of MedleyVox. Audio samples, the dataset, and codes are available on our website.

Index Terms—Multiple singing voices separation, music source separation, speech separation, singing voice separation

1. INTRODUCTION

Singing voice separation (SVS) research has grown remarkably with the progress of deep learning [1] and high-quality benchmark data [2]. The most popular SVS task has been separating singing voices from their accompaniments. Recently, choral music separation [3–7] gained attention, which is the SVS task of extracting individual voices from a choral singing mixture consists of soprano, alto, tenor, and bass parts. However, multiple singing voices separation applicable to popular music remains unstudied.

The two main reasons for the slow advances of multiple singing voices separation are the absence of a benchmark dataset and baseline studies. A benchmark dataset is essential for problem understanding. For example, in speech separation research, it was reported in [8] that recent deep learning-based models have almost achieved the performance of the upper bound. This kind of study can only be performed when there exist high-quality benchmark data [9, 10].

In this paper, we introduce MedleyVox, an evaluation data for multi-singing separation. While building the dataset, we categorize the problem of multiple singing voices separation into four categories. It is done based on various musical properties of singing mixtures, which affects the separation difficulty. This categorization is applied to the items of MedleyVox for detailed performance analysis.

Another challenge for multi-singing separation is that there does not exist any baseline studies including publicly available train data made for such task. To handle the issues, we present a multi-singing mixture construction strategy that leverages various single-singing and speech data for training multi-singing separation models. We also investigate the appropriate analysis/synthesis basis and objective functions for multi-singing separation, and propose the improved super-resolution network (iSRNet) inspired by the recent state-of-the-art speech separation model [11]. Jointly trained with the Conv-TasNet [12], the proposed iSRNet achieves comparable performance to ideal time-frequency masks on duet and unison separation of MedleyVox.

Our work bridges the gap between speech and singing voice separation research. Challenging aspects of multi-singing separation come from the input mixture source consisting of highly correlated vocals. They often have the same pitch with simultaneous on-offset and even consist of segments sung by identical singers, which makes the task more challenging compared to speech separation where speech mixtures rarely consist of one speaker’s different speeches. We provide the baseline studies and carefully analyze these challenges for future research on multi-singing separation.

2. MEDLEYVOX

MedleyVox is an evaluation dataset for multiple singing voices separation. To extend the singing voice separation research towards multi-singing separation, we propose the following four categories that constitute the separation of multiple singing voices. We assume the ideal circumstance where the vocals are cleanly separated from accompaniments by other singing voice separation systems.

i) Unison is the separation where a mixture of two singing voices is given that have identical or octave-shifted melodies, same
3.2. STFT/ISTFT-based Analysis and Synthesis

In speech separation, the current best-performing model [11] on WSJ0-2mix [9] and scale-invariant signal-to-distortion measure is based on the learnable analysis encoder and synthesis decoder given time-domain waveform in/output [12]. However, short-time Fourier transform (STFT) analysis and inverse STFT (iSTFT) synthesis were found effective in universal sound separation [16], where the data distribution is expected to be broader than speech. Similarly, we assume that STFT/iSTFT-based analysis-synthesis framework may outperform the learnable encoder-decoder model in our task because singing shows a broader distribution than speech, e.g., wider fundamental frequency ranges, richer harmonic structures, and longer vocalization. Specifically, we use complex spectral mapping strategy following recent state-of-the-art speech separation [17] and singing voice separation [18, 19] models, which directly estimates real and imaginary coefficients of STFT given an input mixture.

3.3. Objective Functions

Scale-invariant signal-to-distortion ratio (SI-SDR) [20] is a popular training objective in recent speech separation [11, 12]. However, we found that this objective leads to a huge scale of output waveform, which results in a significant clipping error when saving the output in digital format. To handle this, we use a mixture consistency projection during training [21]. Furthermore, we replace SI-SDR with a signal-to-noise ratio (SNR) objective and add a multi-resolution STFT loss [22] to enhance the perceptual quality of network outputs. Since we use a complex spectral mapping strategy for training, we also calculated STFT loss on real-imaginary coefficients of STFT.

3.4. Permutation Invariant Training

In previous singing voice separation or choral music separation, target output source channels are fixed, e.g., vocals and accompaniments [19], or one of four-part harmony [3, 4]. However, in multi-singing separation, the permutation problem [23] exists as it does in speech separation. For example, in duet separation, neural networks have no prior about what output channel should be assigned to a specific voice.

To this end, we used utterance-level permutation invariant training (uPIT) [24] for duet and unison separation where only two vocals exist in input mixture and one-and-rest PIT (OR-PIT) [25] for main vs. rest separation. Especially when using OR-PIT, one major difference between main vs. rest and speech separation [25] is that we can use prior knowledge that the main vocal is louder than the rest vocals in general for popular music. Hence, we intentionally make the main vocal louder than rest vocals during training mixture construction process, thereby no need to calculate objectives for all possible ones and rest assignment; if there are \( N \) single-singing sources in a mixture, the original OR-PIT calculates objectives for \( N \) possible one-and-rest splits and uses the assignment that has the lowest loss for training.
3.5. Improved Super-resolution Network

Inspired by the recent state-of-the-art super-resolution network (SRNet)-based framework on speech separation [11], we propose the improved SR network (iSRNet) for multi-singing separation. As depicted in Fig 2, iSRNet enhances the initial estimates of a backbone separation network, e.g., Conv-TasNet. We replace the original convolutional layers of SRNet with ConvNeXt blocks [26] for efficiency and performance. Also, unlike the original heuristic method [11] of dividing upper and lower frequency ranges roughly by half the Nyquist frequency, we explore the frequency boundaries that separate the frequency regions since we found that Conv-TasNet for multi-singing separation suffers from leakage above 1-5kHz frequency ranges.

4. EXPERIMENTS

4.1. Overview

In Exp 1, we first take an experiment on the appropriate analysis-synthesis frameworks and multi-singing mixture construction strategy (Table 2, 16 kHz sample rate). In Exp 2, we then investigate the effect of mixture consistency projection for preventing output scale exploding (Table 3, 24 kHz sample rate). In Exp 3, we first train the Conv-TasNet-L baseline model, which has an increased channel size from the Conv-TasNet model trained with mixture consistency projection in Exp 2. Then, we improve it by adding SNR, multi-resolution STFT objectives, and using the proposed iSRNet (Table 4, 24 kHz sample rate). In Exp 4, we train main vs. rest separation network (Table 5) with the best-performed setting on duet separation. In Exp 5, ablation studies on the proposed iSRNet are conducted.

For training, we use a total of 400 hours from 13 different single-singing datasets [13, 27–38] and additional 460 hours of LibriSpeech [39] data. Multi-singing mixtures for main vs. rest are constructed using a minimum of 2 to a maximum of 4 singings. For iSRNet, we use the Conv-TasNet-L pre-trained with SNR and multi-STFT objectives for joint training. Detailed hyperparameters are in our GitHub page. For evaluation metrics, both SDR [40] and SI-SDR improvement (SDRi and SI-SDRi) measures are reported in dB.

4.2. Result and Discussion

Exp 1. The results in Table 2 show that on duet separation, STFT-basis slightly outperforms the learnable-basis (Conv-TasNet-learn) in multi-singing separation. Although the evaluation scores are similar, we observe that Conv-TasNet-learn consistently generates crunchy noise unlike Conv-TasNet-STFT. It is also reported in [19] that the waveform-based model for singing voice separation is prone to such artifacts. For subjectively high-quality outputs, we use STFT-basis for the following experiments.

We also observe that our mixture construction strategy – which uses additional speech data and constructing the mixture of the identical singer’s different singings or the same song from different singers – is helpful. Nevertheless, it is noticeable that the model only trained with different singers’ different singings (w/ Only diff singings in Table 2) achieves fair quality contrary to [5], where the random singing mixing for training data resulted in a huge performance decrease. This is perhaps because we used 400 hours of singing data, significantly larger than 104 minutes of [5].

Exp 2. As in Fig 3, without mixture consistency projection during training, the model suffers from output scale exploding due to the scale-invariant nature of SI-SDR objective. This results in significant performance degradation in Table 3 when evaluating the model outputs after save-and-load as 16 bit wave files. Roughly reducing the

Table 2. Evaluation results on different analysis/synthesis basis and the strategy for construction of multiple singing mixtures.

| Model                        | duet  |
|------------------------------|-------|
| w/o Speech data              | 13.3  |
| w/ Only diff singings        | 12.8  |

Table 3. Evaluation on the mixture consistency projection.

| Model                        | duet  |
|------------------------------|-------|
| w/o Mix cons.                | 13.2  |
| w/ save-load                 | 13.2  |
With additional iSRNeting new objectives and networks. As in Table 4, multi-resolution default for the following experiments.

Vocal removal from a multi-singing mixture for a karaoke system.

The output scale is not a solution since keeping the scale of inputs/outputs is relevant for the application usage of singing voice separation, e.g., a main vocal removal from a multi-singing mixture for a karaoke system. To handle the issue, mixture consistency training is used by default for the following experiments.

**Exp 3.** We improve our Conv-TasNet-L baseline model with using new objectives and networks. As in Table 4, multi-resolution STFT loss with SNR objective outperforms the baseline models. With additional iSRNet, our model achieves 1 dB of performance gain compared to the baseline on SI-SDRi.

On unison separation also in Table 4, our model shows comparable performance to ideal masks and even outperforms ideal ratio mask (IRM) on SDRi. Comparing the performance of the ideal binary mask (IBM) and IRM on duet and unison, we notice that the scores on unison is more than 10 dB lower than the scores on duet. This implies that phase information is critical for unison separation. Unlike our model which is trained to work on both duet and unison separation, a model that is trained solely for unison with cIRM objective (e.g., [41]) can be an option to maximize the performance in unison, but it is out of the scope of this paper.

**Exp 4.** We validate main vs. rest separation model with the best-performing setting in duet separation, which is, SNR loss combined with multi-resolution STFT loss (Conv-TasNet-L in Table 5) and the additional iSRNet (w/ iSRNet in Table 5). Similar to duet separation, the proposed iSRNet brings performance gain compared to the baseline model. However, there is a significant performance gap between the oracle methods (e.g., SI-SDRi 13.2 dB on IRM) and the proposed iSRNet (6.3 dB). We will discuss this failure in a later section.

**Exp 5.** In ablation studies on the proposed iSRNet as in Table 6, by simply replacing the original convolutional layers of SRNet into ConvNeXt blocks, the performance rises 0.5dB with 40 times smaller number of trainable parameters and computational costs. Also, the model achieves the best performance when using a heuristic frequency upper boundary at 3 kHz.

**Analysis of Failure Cases.** We have analyzed several failure examples that have significantly lower scores than averages. On duet separation, we confirm that some examples suffer from sudden singer assignment changes; a singer of the estimation A is suddenly assigned to the estimation B at specific regions. We also find some cases that model tends to output worse estimations when the input examples have many silenced regions. We suspect this behavior comes from the usage of a global layer normalization layer [12], which normalizes whole given sequences without separately considering silenced regions.

In main vs. rest separation, the model frequently suffers from the circumstance that the section originally corresponds to silence regions in a main vocal part wrongly filled with one of the rest vocals. This is mainly because we train the model using OR-PIT with the input mixture in which the main vocal is louder than the rest, whereas silence regions are quieter than the rest vocals. Also, the model suffers from extracting an unintended vocal for the main. This is also because of our OR-PIT training scheme; the main vocal is louder for most popular music; however, not always. We consider that the approach of separating every source in a mixture at once [42] could be the solution for these issues, which is also directly applicable to N-singing separation.

**Future works.** Since we performed the experiments on segment-wise, long input sequences such as 3-4 minutes songs have to be considered in future works for an application usage. Unfortunately, plain chunk-wise processing that has been frequently used in continuous speech separation [43, 44], which assigns the output speakers based on the similarity of overlapped regions is not feasible to music because the multi-singing mixtures possibly contain much longer silences unlike speeches; appropriate chunk window and hop sizes would be a solution but it seems impossible to find a universal setting that perfectly fits every song.

Chunk-wise processing that is robust on long silences proposed in [45] can be another option to handle this. However, the question remains because such method used speaker information for processing and assignments. Since multi-singing mixtures are often composed with identical singer’s singings, assignments based on singer representations can lead to sub-optimal results. To overcome this, it is necessary to study the robust representations for singer assignments that reflect not only speaker-related information such as timbre, but also context-related information such as melodic sequences.

**5. CONCLUSIONS**

In this paper, we introduced MedleyVox, an evaluation dataset for separation of multiple singing voices into each constituting singings. We presented a multi-singing mixture construction strategy that uses various single-singing and speech data to overcome the absence of publicly available training data for multi-singing separation networks. Also, the proposed iSRNet achieved superior performance on duet and unison separation when jointly trained with the Conv-TasNet. Lastly, we thoroughly analyzed the remaining challenges for multi-singing separation.

### Table 4. Evaluation results on investigating optimal objectives, the proposed iSRNet, and ideal time-frequency masks on duet and unison separation.

| Model          | duet | unison |
|----------------|------|--------|
|                | SDRi | SI-SDRi | SDRi | SI-SDRi |
| Conv-TasNet-L  | 14.2 | 13.2 | 14.5 | 14.1 |
| w/ SNR loss    | 13.9 | 12.9 | 14.4 | 14.0 |
| w/ Multi-STFT loss | 14.4 | 13.4 | 14.5 | 14.0 |
| w/ iSRNet      | **15.1** | **14.2** | **14.9** | **14.4** |
| IBM            | 16.5 | 15.9 | 15.5 | 15.1 |
| IRM            | 16.1 | 15.3 | 15.4 | 15.0 |
| cIRM           | 56.3 | 56.5 | 51.5 | 51.9 |

**Table 5.** Evaluation results of the baseline, the proposed iSRNet, and ideal time-frequency masks on main vs. rest separation.

| Model          | main vs. rest |
|----------------|--------------|
|                | SDRi | SI-SDRi |
| Conv-TasNet-L  | 7.0  | 5.8   |
| w/ iSRNet      | 7.2  | 6.3   |
| IBM            | 14.2 | 13.7  |
| IRM            | 13.7 | 13.2  |
| cIRM           | 56.9 | 57.0  |

**Table 6.** Ablation studies on the proposed iSRNet with various heuristic frequency upper boundaries.

| Model | Freq. boundary | duet | unison |
|-------|---------------|------|--------|
|       | SDRi | SI-SDRi | # of | MACs |
| SRNet | 6k   | 14.4 | 13.5  | 6.67M | 1.93T |
| iSRNet| 1.5k | 14.7 | 13.8  | 166.1K | 47.5G |
6. REFERENCES

[1] Y. Mitsufuji, G. Fabbro, S. Uhlisch, and F.-R. Stöter, “Music demixing challenge 2021,” arXiv:2108.13559, 2021.
[2] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bitner, “Msysdb18-hq - an un compressed version of msysdb18,” doi.org/10.5281/zenodo.3338373, 2019.
[3] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, “Deep learning based source separation applied to choir ensembles,” arXiv:2008.07645, 2020.
[4] M. Gover and P. Depalle, “Score-informed source separation of choral music,” in Proc. of ISMIR, 2020.
[5] S. Sarkar, E. Benetos, and M. Sandler, “Vocal harmony separation using time-domain neural networks,” in Proc. of InterSpeech, 2021.
[6] K. Schulze-Forster, C. SJ Doire, G. Richard, and R. Badeau, “Unsupervised audio source separation using differentiable parametric source models,” arXiv:2201.09592, 2022.
[7] K. Chen, H.-W. Dong, Y. Luo, J. McAuley, T. Berg-Kirkpatrick, M. Puckette, and S. Dubnov, “Improving choral music separation through expressive synthesized data from sampled instruments,” arXiv:2209.02871, 2022.
[8] S. Lutati, E. Nachmani, and L. Wolf, “Sepit approaching a single channel speech separation bound,” in Proc. of InterSpeech, 2022.
[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. of ICASSP, IEEE, 2016.
[10] A. Défossez, “Hybrid spectrogram and waveform source separation approach using neural networks with multi-resolution spectrogram,” in Proc. of ICASSP. IEEE, 2020.
[11] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multitalker speech separation,” in Proc. of ICASSP. IEEE, 2017.
[12] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” TASLP, 2017.
[13] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, “Continuous speech separation: Dataset and analysis,” in Proc. of ICASSP, IEEE, 2020.
[14] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeberlin, “All-neural online source separation, counting, and diarization for meeting analysis,” in Proc. of ICASSP. IEEE, 2019.