EXPLORING ALIGNED LYRICS-INFORMED SINGING VOICE SEPARATION

Chang-Bin Jeon, Hyeong-Seok Choi and Kyogu Lee
Department of Intelligence and Information
Music and Audio Research Group (MARG)
Center for Superintelligence
Seoul National University
{vinyne, kekepal5, kglee}@snu.ac.kr

ABSTRACT
In this paper, we propose a method of utilizing aligned lyrics as additional information to improve the performance of singing voice separation. We have combined the highway network-based lyrics encoder into Open-unmix separation network and show that the model trained with the aligned lyrics indeed results in a better performance than the model that was not informed. The question now remains whether the increase of performance is actually due to the phonetic contents that lie in the informed aligned lyrics or not. To this end, we investigated the source of performance increase in multifaceted ways by observing the change of performance when incorrect lyrics were given to the model. Experiment results show that the model can use not only just vocal activity information but also the phonetic contents from the aligned lyrics.

1. INTRODUCTION
Singing voice separation is one of the most widely studied areas in the field of audio signal processing. In particular, the importance of research is greatly emphasized because it can contribute to the pre-processing step of research in various fields of Music Information Retrieval (MIR), such as automatic music transcriptions and automatic lyrics alignments. With the recent development of deep neural networks, a number of music source separation studies have been published and showed excellent performance. These studies have a common feature of separating music source by using only information from the sound source itself, such as a 1-dimensional waveform [5, 19] or a 2-dimensional spectrogram [9, 20, 23].

One of the distinguishing characteristics that differentiate music signals from other audio signals is that usually there exists the corresponding music scores or lyrics. Therefore, several studies attempted to separate the sources by utilizing additional information other than the information of the sound source itself. For example, the additional information such as pitch [1] or even the whole score [26] can be used as a prior to help separate the source of interest from the mixture. In general, however, music scores for certain songs are not readily available, while lyrics can be easily collected on the web.

The lyrics are particularly closely related information to singing voice; thus, have promising possibility to be used as additional information for singing voice separation. Recent studies proposed the ways of using linguistic features extracted from the end-to-end automatic speech recognition model [24] or voice conversion model [3] to the singing voice separation framework. However, the way of using explicit lyrics information has not been studied enough so far, which motivates us to study the possibility of lyrics-informed singing voice separation. We expect that singing voice separation systems can benefit from lyrics information because of the rich information contained in the phonetic features such as formant frequencies.

To utilize the lyrics, we combined the highway network-based lyrics encoder [11] into the current state-of-the-art music source separation network, Open-unmix [20]. In addition, we tried two conditioning methods — 1. local conditioning, 2. concatenation — and compare the performance. Note that the alignment between the lyrics and songs itself is another separate line of research [6, 18]. For our study, we assume that the alignment is already done and only focus on the use of the aligned lyrics.

The information in the aligned lyrics can be seen from two perspectives: 1. The timing information of vocal activity, 2. phonetic information. Therefore, it is important to check if the network is using the phonetic information other than the vocal activity information. Various evaluations were conducted to examine whether the phonetic information of the aligned lyrics actually contribute to improving the performance. We found that the proposed model trained with the aligned lyrics clearly show better performance than the baseline model trained without any additional information. Furthermore, the experiment results show that the performance of the proposed model even exceeds the model trained only with additional vocal activity information. To the best of our knowledge, this is the first research to directly use the lyrics information for a singing voice separation task.
2. RELATED WORK

2.1 Informed Source Separation

Several studies using machine learning algorithms other than deep neural networks attempted to separate singing voice with side information. For example, robust Principal Component Analysis (rPCA) was used with additional vocal activity information [2]. Also, rPCA and Non-negative Matrix Factorization (NMF) were used with pronounced lyrics [4].

Although only few studies have tried to use additional information for singing voice separation using deep neural networks, it was reported that vocal activity information can be used as an input to the network along with spectrogram to enhance the performance of the singing voice separation network [14]. Very recently, in the speech enhancement field, attempts have been made to utilize text information to increase the separation performance [15].

2.2 Open-unmix

Open-unmix [20] is the state-of-the-art music source separation network using the MUSDB18 dataset [13]. In particular, it consists of 3 bi-directional Long Short-Term Memory (LSTM) layers for source separation with 3 additional fully-connected layers. Batch normalization [8] was used after every fully-connected layer and skip connection [7] was used between the inputs and outputs of 3 consecutive bi-directional LSTM layers. Trainable input and output scalers through frequency-axis are also the special features of Open-unmix, differentiating from other studies that use decibel scales.

We used Open-unmix network as our baseline model because we wanted to check whether the aligned lyrics information could improve the performance of the current state-of-the-art model. The channel inputs and outputs are mono in our study although the original study used stereo. This is because our singing dataset is made up of a clean singing voice without any reverberation, chorus, or doubling, as opposed to the singing tracks in MUSDB18’s singing voice dataset, which are already processed for the stereo. Details and the full structure of the baseline Open-unmix are illustrated in Figure 1.

2.3 Singing Voice Synthesis

Singing voice synthesis and lyrics-informed singing voice separation tasks share a similar framework in that the input and output are the same as lyrics and singing voice spectrograms, respectively. Recently, [11] proposed the singing voice synthesis network that succeeds in creating high-quality singing based on 60 Korean songs sung by a single singer. It is based on the Text-to-Speech model [22], which consists of 1-dimensional convolutional neural networks and highway networks [17]. Therefore, we borrowed the idea of using the highway network-based lyrics encoder and integrated it into the source separation network.

3. PROPOSED METHOD

3.1 Lyrics Encoder

The detailed structure of the lyrics encoder is shown in Figure 2 and the structure of the highway networks used in the lyrics encoder is defined as follows,

\[ y = \text{ReLU}(x \ast W_H) \cdot \sigma(x \ast W_T) + x \cdot (1 - \sigma(x \ast W_T)), \quad (1) \]

where \( x \) and \( y \) are input and output of the network and \( \cdot \) refers to the element-wise multiplication. \( x \ast W_H \) and \( x \ast W_T \) are the 1-dimensional convolution layers which have the same input and output channel sizes. Biases are omitted in Eqn (1). Zero-padding was applied to keep the
input and output length the same in every convolution layer. Dropouts [16] with 0.05 dropout rate, were applied after the activation functions. Dilated convolution [27] were used to expand the receptive field to 165 frames. It is about seven times larger than the model without the dilation. In our experimental settings, 165 frames are equal to about 1.915 seconds.

3.2 Local Conditioning of Lyrics Encoding

The local conditioning method [11] was used to insert the encoded lyrics information into the singing voice separation network. The local conditioning is defined as follows,

\[ y = \text{ReLU}(x \ast W_f) \cdot \sigma(x \ast W_g + L_2), \]

where, \( x \ast W_f \) and \( x \ast W_g \) are the 1-dimensional convolution layers which have same input and output channel sizes. \( L_1 \) and \( L_2 \) are equally separated features through the channel axis from the output of the consecutive lyrics encoder and the 1-dimensional convolutional layer. The 1-dimensional convolution layer with filter size 1 was added on the output of the lyrics encoder so that the channel size of each \( L_1 \) and \( L_2 \) could be 512, \( \sigma \) refers to the sigmoid activation function. Details of the full structure are in Figure 3a.

3.3 Concatenation of Lyrics Encoding

A concatenation method, which is a simple but powerful conditioning method, was also used in our study to pass the encoded lyrics information into the singing voice separation network. By concatenating the output of the first fully-connected layer and the lyrics encoder output, the channel size of LSTM layers input becomes 1024. Also, the channel size of the second fully-connected layer input becomes 1536 by the skip-connection of LSTM input and output. Details of the full structure is in Figure 3b.

Figure 3: The structure of the Open-unmix networks combined with the lyrics encoder using (a) local conditioning and (b) concatenation method.

Table 1: The composition of our singing dataset.

| Singer | Gender | Train | Validation | Test | Total |
|--------|--------|-------|------------|------|-------|
| 1      | Female | 79    | 5          | 8    | 92    |
| 2      | Male   | 8     | 2          | 0    | 10    |
| 3      | Female | 8     | 2          | 0    | 10    |
| 4      | Female | 9     | 0          | 1    | 10    |
| 5      | Female | 8     | 0          | 1    | 9     |
| 6      | Male   | 10    | 0          | 0    | 10    |
| 7      | Female | 8     | 0          | 2    | 10    |
| 8      | Female | 9     | 1          | 0    | 10    |
| 9      | Male   | 9     | 0          | 1    | 10    |
| 10     | Male   | 7     | 1          | 1    | 9     |
| 11     | Female | 7     | 3          | 0    | 10    |
| 12     | Female | 0     | 5          | 5    | 10    |
| 13     | Male   | 0     | 0          | 1    | 1     |

4. EXPERIMENTS

4.1 Dataset

Here we used a total of 201 Korean pop songs sung by 13 amateur singers as target clean singing sources. This dataset has a total length of 11 hours and 44 minutes. Of these, we used 162 songs (9h 3m) for training, 19 songs (1h 7m) for validation, and 20 songs (1h 7m) for test dataset. The detailed composition of the dataset is described in Table I.

We aligned Korean syllable following [11]; one Korean syllable is made up of onset (consonant), nucleus (vowel) and coda (consonant), we aligned onset and coda for 4 frames, and nucleus for other frames. The example of the alignment applied to the spectrogram is shown in Figure 6.

A total of 19,113 instrumental songs were used as accompaniment for training networks because we did not have real accompaniment tracks corresponding to the singing voice dataset. Since various studies using MUSDB18 [13] or DSD100 [12] database also used random mixing techniques, i.e. creating random accompaniment for each iteration that was not related to the original singing, we decided that using arbitrary accompaniments would not be a problem for training. In addition, if we used the same specific accompaniments for the test singing voice dataset, we assumed that it is reasonable for identifying how the information containing the phonetic features of the aligned lyrics has changed the performance of the network, the fact we wanted to identify. Therefore, for the validation and test dataset, we randomly chose each 19 and 20 instrumental songs which have a longer length than the singing data, and shorten the length to the same with the singing.

4.2 Training

In our singing dataset, the number of songs recorded by the first singer is outnumbered compared to the others, accounting for about 46 percent of the total. In order to prevent bias to a particular singer when training the networks, a singer was first selected with the same probability when constructing a batch to be used for each iteration, and the vocal source to be used for training was sampled only for the songs recorded by the selected singer.
Figure 4: The example of inserting the aligned lyrics to the networks.

A mono sound source with a sample rate of 22050 Hz was used in the experiment. FFT point size and window size were set to 1024 samples (0.0464 seconds) to convert them into a spectrogram, and Short-time Fourier transform (STFT) hop size to 256 samples (0.0116 seconds). Adam optimization method [10] was used for training with a learning rate of 0.001, $\beta_1$ for 0.9, $\beta_2$ for 0.999. Mean squared error (MSE) loss function between the ground-truth and the outputs of the models were used in our study. We trained the models for 500 epochs with calculating validation loss for every epoch. The learning rate was reduced to 30 percent if there was no decrease in validation loss during 25 epochs. Early stopping was applied after 50 epochs without a decrease in validation loss.

4.3 Evaluation Methods

Here we briefly summarize the various experiments that will be shown in the following Section 5. The experiments will be conducted in three following ways.

First, in Section 5.1, we compare and analyze how much performance improvement there are between the baseline model trained without the lyrics and the model trained with the aligned lyrics.

Second, in Section 5.2, we check if the network exploits the vocal activity information included in the aligned lyrics. The lyrics include both the vocal activity information and phonetic information, and thus expected to use the vocal activity information correctly.

Third, in Section 5.3, we check if the given input is correctly being used by the network trained with aligned lyrics. This experiment was done by giving incorrect inputs in the evaluation stage. It is expected that the incorrect inputs will significantly reduce performance.

For network performance evaluations, Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) scores [25] were computed by museval python library [21].

5. RESULTS

The configuration of the models we trained in our experiments is in Table 2. We trained four models each using local conditioning and concatenation method. model 1 is the baseline model that trained without the lyrics encoder. model 2 is the model that trained with meaningless 0 value inputs to the lyrics encoder. This model is only for checking the performance change of the networks caused by the extended network capacity. model 3 is the model that trained with only vocal activity information to the lyrics encoder. We simply used 0 value as unvoiced sections and 1 as voiced sections so that the 128-dimensional embedding can train useful meaning from it. model 4 is that trained with aligned lyrics information. For both model 3 and model 4, note that 0 value has clear meaning, unvoiced sections, unlike 0 value in model 2 is meaningless. Since the baseline model is the same for each local conditioning and concatenation method, we trained a total of 7 models for the experiments. Except model 1, we will use the abbreviation of local conditioning and concatenation methods, each LC and CC, in front of the model names for convenience. For example, the model trained with aligned lyrics and the local conditioning method is LC-model 4.

5.1 Performance Evaluation

The quantitative performance evaluation scores of the models are shown in Table 3. Median scores were taken from the median values of 20 songs, which were calculated for every frame (a median of frames, a median of tracks). Mean scores represent the scores taken by a mean of frames, a mean of tracks. Each frame was set to 1 sec-
| Models     | Median   | Mean   |
|-----------|----------|--------|
|           | SDR      | SIR    | SAR    | SDR      | SIR    | SAR    |
| model 1   | 9.956    | 18.674 | 9.847  | 8.595    | 16.062 | 9.145  |
| LC-model 2| 10.140   | 18.465 | 9.766  | 8.589    | 16.001 | 9.093  |
| LC-model 3| 10.090   | 18.713 | 9.763  | 9.250    | 16.298 | 9.153  |
| LC-model 4| 10.767   | 19.505 | 10.223 | 9.723    | 17.116 | 9.699  |
| CC-model 2| 10.110   | 18.434 | 9.909  | 8.691    | 16.164 | 9.207  |
| CC-model 3| 10.444   | 19.328 | 10.169 | 9.718    | 17.031 | 9.609  |
| CC-model 4| 10.757   | 19.623 | 10.371 | 9.752    | 17.250 | 9.803  |

Table 3: Evaluation scores of our singing voice separation models. All scores are in [dB] scale.

It was confirmed that the separation performance of both LC-model 4 and CC-model 4 improved from the model 1. This implies that aligned lyrics information can be used as helpful features for singing voice separation networks. Comparing to LC-model 3 and CC-model 3, we could verify that there were clear performance gains not only from the lyrics alignment information but also the phonetic features of the lyrics itself. It was also confirmed that the performance gains do not come from just network capacity growth, given that there are no significant differences in the performance of model 1 and model 2. The spectrograms of the separated sample are in Figure 5.

Despite the expectation that the vocal activity information is powerful information to the networks, performance gains observed in LC-model 3 were very slight. It was much smaller than the improvements achieved from CC-model 3. By these, we analyzed that the concatenation method is slightly better for making the networks to reflect the vocal activity information. Nevertheless, we considered that both conditioning methods were effective when giving the networks aligned lyrics information.

5.2 Analysis of Vocal Activity Information Usage

In this section, we quantitatively assessed how well the networks leverage vocal activity information of aligned lyrics. The purpose of this is to see if the models have not been trained by focusing only on either one of the vocal activity information or the phonetic information of aligned lyrics, which are both critical for the separation performance.

To this end, the separated spectrogram values were divided by the largest values of each source for normalization, so that the minimum and maximum values become 0 and 1. Then, the energy of each time axis was summed to create a vector that contains vocal activity information. It was decided whether vocal activity exists or not based on whether the values were larger or smaller than 0.1 as was done in [4]. Precision, recall, and F1 scores were calculated with the created vocal activity vectors by taking the place where the lyric exists as the ground-truth voiced sections. Scores are contained in Table 4.

From the results of Table 4, we have confirmed that model 4 can separate the vocal source by reflecting the vocal’s timing information more accurately than model 1 for both lyrics conditioning methods. Also, it was confirmed that model 3 achieved higher scores for all measures than model 4. This is a reasonable result because model 3 were trained with vocal activity information only, while model 4 needed to learn how to leverage both vocal activity information and phonetic information appropriately while training. Nevertheless, F1 score differences were negligible, which means model 4 was also capable of reflecting timing information as well as model 3.

5.3 Analysis with Using Incorrect Lyrics

To check if model 4 effectively uses the information in lyrics, we observed the performance change when incorrect lyrics were given as input during the evaluation stage. The results are shown in Table 5.

If the networks had learned to effectively use the information in lyrics it is expected to output silence when the lyrics meaning unvoiced sections are given in the evaluation step. To validate this assumption, we inserted 0 values (Zero) to the lyrics encoder in to model 4 in the evaluation step. As expected, almost every sound has been erased from the mixture with only a little noise left as seen in Figure 7 and critical performance degradation, over 10 dB in SDR score, has occurred.

Furthermore, performance degradation was observed when all the lyrics were replaced with random values
### Table 4: The precision, recall, and F1 scores to evaluate how well the networks used the vocal activity information from aligned lyrics.

| Models          | Precision | Recall | F1 score |
|-----------------|-----------|--------|----------|
| model 1         | 0.807     | 0.853  | 0.828    |
| LC-model 2      | 0.810     | 0.852  | 0.830    |
| LC-model 3      | 0.887     | **0.857** | 0.872   |
| LC-model 4      | 0.876     | 0.854  | 0.865    |
| CC-model 2      | 0.814     | 0.853  | 0.833    |
| CC-model 3      | **0.896** | 0.855  | **0.875** |
| CC-model 4      | 0.879     | 0.855  | 0.867    |

### Table 5: Performance comparisons when different inputs are given in the evaluation stage. Zero: 0 value inputs. Random: Random value inputs. VA+Random: Replace voiced sections with random value. AlignedLyrics: Aligned lyrics (The proposed method). All scores are in [dB] scale.

| Models          | Inputs      | SDR   | SIR  | SAR  |
|-----------------|-------------|-------|------|------|
| LC-model 4      | Zero        | 0.001 | 8.040| -4.286|
|                 | Random      | 5.317 | 15.802| 4.845|
|                 | VA+Random   | 7.899 | 19.270| 6.403|
|                 | AlignedLyrics | **10.767** | **19.305** | **10.224** |
| CC-model 4      | Zero        | 0.002 | 7.246|-3.671|
|                 | Random      | 0.946 | 14.837| 0.341|
|                 | VA+Random   | 7.164 | 19.545| 6.290|
|                 | AlignedLyrics | 10.757 | 19.623| 10.371|

Figure 7: The examples of the separated vocal spectrograms with incorrect inputs and correct aligned lyrics inputs were given to LC-model 4. The dashed line shows the enhanced parts when the aligned lyrics are used. Note that the region is closely related to the formant frequencies.

### 6. CONCLUSION

In this study, we proposed an integrated framework of combining the lyrics encoder into the state-of-the-art Open-unmix separation network. Local conditioning and concatenation methods were shown to be able to effectively condition the aligned lyrics into the singing voice separation networks. Through various experiments, it was confirmed that the phonetic information of aligned lyrics can contribute to the performance improvements as well as the vocal activity information. We plan to use the unaligned lyrics for the singing voice separation for the future works.

### 7. ACKNOWLEDGEMENTS

This work was supported partly by Kakao and Kakao Brain corporations and partly by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7078548).

### 8. REFERENCES

[1] Estefanía Cano, Gerald Schuller, and Christian Dittmar. Pitch-informed solo and accompaniment separation towards its use in music education applications. *EURASIP Journal on Advances in Signal Processing*, 2014(1):23, 2014.

[2] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang. Vocal activity informed singing voice separation...
with the ikala dataset. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 718–722. IEEE, 2015.

[3] Pritish Chandna, Merlijn Blauw, Jordi Bonada, and Emilia Gómez. Content based singing voice extraction from a musical mixture. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 781–785. IEEE, 2020.

[4] Z Chen, PS Huang, and YH Yang. Spoken lyrics informed singing voice separation. In Proc. HAMR, 2013.

[5] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254, 2019.

[6] Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. Acoustic modeling for automatic lyrics-to-audio alignment. arXiv preprint arXiv:1906.10369, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[9] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[11] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee. Adversarially trained end-to-end korean singing voice synthesis system. arXiv preprint arXiv:1908.01919, 2019.

[12] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In International conference on latent variable analysis and signal separation, pages 323–332. Springer, 2017.

[13] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017.

[14] Kilian Schulze-Forster, Clément Doire, Gaël Richard, and Roland Badeau. Weakly informed audio source separation. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 273–277. IEEE, 2019.

[15] Kilian Schulze-Forster, Clement SJ Doire, Gaël Richard, and Roland Badeau. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7274–7278. IEEE, 2020.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958, 2014.

[17] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. arXiv preprint arXiv:1505.00387, 2015.

[18] Daniel Stoller, Simon Durand, and Sebastian Ewert. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 181–185. IEEE, 2019.

[19] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185, 2018.

[20] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-umix - a reference implementation for music source separation. Journal of Open Source Software, 2019.

[21] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK, pages 293–305, 2018.

[22] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4784–4788. IEEE, 2018.

[23] Naoya Takahashi, Naobarun Goswami, and Yuki Mitsufuji. Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pages 106–110. IEEE, 2018.

[24] Naoya Takahashi, Mayank Kumar Singh, Saky Basak, Parthasarathy Sudarsanam, Sriram Ganapathy, and Yuki Mitsufuji. Improving voice separation by incorporating end-to-end speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 41–45. IEEE, 2020.
[25] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[26] John F Woodruff, Bryan Pardo, and Roger B Dannenberg. Remixing stereo music with score-informed source separation. In *ISMIR*, pages 314–319, 2006.

[27] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.