Mel2Seq-VS: Vocal Separator for Pop Music using Mel Filter banks and Generative Adversarial Nets

Hao-Yuan Tang
Department of Electrical and Computer Engineering
University of California, San Diego
h6tang@ucsd.edu

Abstract

Source separation is a task to extract specific information from raw audio sequences. However, compared to speech signals, the sources contained in music signals are more complicated because of factors like variant instruments and strong beats in the backgrounds. One common idea for source separation is utilizing mel filter bank and generative adversarial nets(GAN). In this project, instead of extracting target channels from pop music audios, we aim to develop a deep learning workflow using mel filter bank and GANs to synthesize source of the target. Compared to common frameworks utilizing spectrogram, we tried to explore the usage of mel spectrogram-GAN based models using transfer learning strategy. To best of our knowledge, this project is the first research to study source separation using mel spectrogram-GAN based models for pop music.

1. Introduction

1.1. Problem Statements

When producing music, recordings of individual instruments were often mixed into single audio signal to increase variance of sound effects. For better hearing experience, music workers used source enhancement to increase the quality of single channel from the music track, especially for pop music with vocals by singers. Music source separation(MSS) is an essential start step in source enhancement workflow [2, 5, 16]. After retrieving the target source, then some following post-processing techniques could be apply to increase audio perceptual quality.

However, compared to speech signal, music tracks have much more complicated scenarios for extracting one target source. The first challenge is variant instruments in backgrounds. Each instrument has its corresponding frequency range and tempos [1], which means data distribution is more difficult to be learned in time and frequency domain across the dataset. The problem by instruments and vocals for MSS task is not only the multi-source separation problem, but also the interactions between each of them. For example, many singers used to intentionally stretches their voiced sounds to match other musical instruments which makes vocals more indistinguishable. Another problem is the ”beats” within pop music. Some low-frequency range beats cover the whole music tracks and sometimes their energy is even stronger than the target source. Therefore, models for this application must have ability of handling different frequency bands of the instruments.

The second challenge is the large variance of musical factors across the relevant dataset, like MUSDB18 [18]. These factors significantly affect performance of MSS task, such as genres [25], singer identity [20], style [19], and so on. For example, intuitively, extracting single source from slow tempo and relaxing music, such as country music, would be quite different from genre like heavy metal, which usually contains strong beats and growling/screaming style vocals. Data distributions between different genres or styles make source separation more challenging for machine learning workflows. After experiments, some statistical methods using Gaussian distribution assumption were failed for the MSS tasks.

Another challenge comes from wide dynamic range in music tracks. In general, speech occurs at lower overall intensity levels than singing which features a wider variation in intensities. In contrast, differences in dynamic ranges of music tracks, including frequency bands and intensities, would set certain limitations for singing voice detection for feature-based classification methods [9, 27, 29]. Therefore, the candidate models in this project must be able to handle wide dynamic ranges in time and frequency domain.

1.2. Previous Approaches

Here we discussed common and popular approaches **audio source separation** for both speech and music applications because there are relatively more mature and well-developed models in speech tasks compared to music. They shared similar idea when designing source separation work-

flows. For audio source separation, the most popular methods could be categorized into 3 classes depending on the used musical features: **sequence-based**, **image-based** and **hybrid** of the previous two types.

For **sequence-based** methods, the input data are processed or raw sequences of source audios [10, 19, 22, 23]. The time dependency for both speech and music is crucial for source separation. To capture this dependency, some studies developed models which take recurrent neural network (RNN) [24] or 1D convolutional neural network [22] as backbone architectures. However, the limitations of receptive fields in CNN and short-term dependency in RNN sometimes set constraints on model's ability of capturing long-term dependency for longer sequence [24]. Therefore, a common solution is to split these input into small chunks (like 25ms) for models, but this is computational inefficient when reconstructing a long signal. Also, the lack usage of frequency domain features is unfavorable for dealing instruments problem in music application. So, the models to use in this application should utilize features in both of time and frequency domains and process longer input sequences.

Alternatively, for **image-based** models, the inputs are usually transformations of the whole sequences. The most common workflow in this case is firstly using short time Fourier transform (STFT), which produces the corresponding spectrogram of input signals, then used CNN-based model for reconstruction. Since spectrogram is 2D arrays, these types of works treat source separation as an "image" separation problem [5]. In other words, either to predict a mask [**?**, 3] for filtering target source out in time-frequency domain or to generate another target spectrogram [14, 24]. However, the limited receptive field by CNN is still unsolved, and the dimension after STFT sometimes could be very large which is bad for computation and resource usage.

The usage of **mel filter bank** is getting popular and has widely gained successes in audio processing applications [6, 7, 17, 21, 28]. In short, the idea behind mel filter bank is to mimic how human ears percept sounds, and this technique can preserve condensed time-frequency domain features in low dimensional feature space. The details and advantages of mel filter bank will be discussed in next section. However, one drawback of mel filter bank is the "inversion" problem. Mel spectrogram by mel filter bank, unlike the relationship between STFT and inverse-STFT (iSTFT), the inversion to sequence can only achieved via approximation, such as Griffin-Lim algorithm [15], which always companies with residual errors, as mentioned in documents of Python package *librosa* [11].

To overcome the inversion problem, solution for source separation or synthesis using mel filter bank were designed as **hybrid** models of "sequence-based" and "image-based". Instead of approximation, it's more "simple" to **generate** reconstruct the target source. More specifically, the most popular models took mel spectrogram as the input and output a corresponding reconstructed signal using Generative Adversarial Nets (GANs) [6, 7, 17, 21, 28]. In this way, mel-scaled condensed features in time-frequency domain could be utilized at the same time and the inversion of mel spectrogram could be handled by the generators under an adversarial training strategy.

1.3. Research Goal

In this project, firstly, we want to investigate whether mel filter bank could also be helpful for MSS task, especially for pop music tracks. Secondly, since most of these well-developed models were designed for speech applications, such as speech synthesis [6, 7, 17, 28] and text-to-speech [21], we want to investigate the potential usage of mel filter bank by transfer learning from these models. The research target is transferred from speech to pop music, so the complicated scenario and factors could make the main task more challenging. Finally, we aim to build a real application: vocal separator, to demo the potential values of mel filter bank in pop music industry.

2. Methods

2.1. Dataset

The dataset in this project was scrapped from online streaming platforms. There are 30 different pop music tracks from different Billboard top albums with variant singers and genres, including country, choir, rock, rap, hip hop, and so on. To handle dynamic ranges of intensities, all tracks were normalized in loudness following standard ATSC A/85 and recorded in monoaural with sampling rate 22050 Hz and 16000 Hz for a model called HiFi-GAN-VC [26]. For better utilization of GPU memory, all tracks were split into around 90 6-second chunks. There's no other preprocessing step for the dataset.

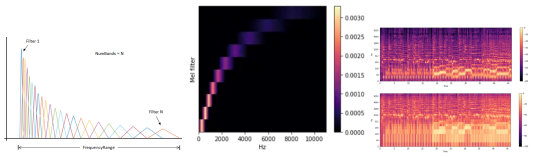
2.2. Mel Filter Bank

Mel filter bank was designed as half-overlapped triangular filters equally spaced on the mel scale to the power spectrum to extract frequency bands. The mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies. We can convert between Hertz (f) and mel (m) using the following equations:

$$m = 2595 \log(1 + \frac{f}{700}) \tag{1}$$

$$f = 700(10^{10m/2595} - 1) \tag{2}$$

Each filter in the filter bank is triangular having a response of 1 at the center frequency and decrease linearly towards 0 till it reaches the center frequencies of the two



(a) (b) (c)

Figure 1. Mel filter bank (a) filter bank on mel-scaled (b) visualization of filter bank(n_mel = 80) (c) Spectrogram by STFT(up) and corresponding mel spectrogram(n_mel = 80)

Parameter	Value
num_mels	80
n_fft	1024
window size	1024
hop length	256
mel_min	0.0
mel_max	8000.0

Table 1. Parameters in mel filter bank

adjacent filters where the response is 0 , as shown in this Figure 1-a. After passing through mel filter bank, as we can observe in Figure 1-c, the distribution of mel spectrogram in frequency domain is quite different from STFT spectrogram: with more energy distributed in low frequency bands and the dimension in frequency axis were reduced from 8192 to 80 which is the **number of filters** in mel filter bank. We expected this behavior to be useful when we handle multi-band instruments conditions in pop music. Intuitively, the energy of instruments with high frequency ranges could be scaled to low frequency ranges for condensation.

In implementations, mel filter banks were constructed in consistent with authors of the proposed models. The first step is to construct "triangular-liked" filter basis using Hanning windows equally distributed in the mel-scale. Then we passes signals through the filter bank and get mel spectrogram for training and inference input. More parameters in the filter bank were shown in Table 1.

2.3. Models

As mentioned in section 1.3, the most popular models using mel filter bank were designed for speech applications, so the usage of these models on MSS task is what we have interests. The candidate models are MelGAN [7], HiFi-GAN [6], waveglow [17], HiFi-GAN for voice conversion (HiFi-GAN-VC) [26]. These models were designed as generative adversarial nets whose generator takes mel spectrogram as input and output a generated audios. In this case, the generated audio should only contained vocals of singers and the loss of generator were calculated between generated and real vocals. For discriminator, the training task is to identify whether the generated vocals were real or fake.

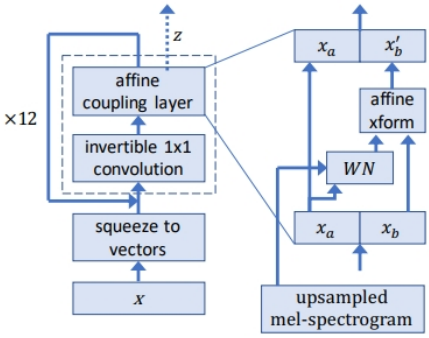


Figure 2. WaveGlow

More details of models will be discussed as follows.

2.3.1 WaveGlow

WaveGlow [17] was proposed by NVIDIA corporation for speech synthesis. WaveGlow is a very high capacity generative flow consisting of coupling and invertible convolutions, with each coupling layer consisting of a stack of layers of dilated convolutions. In short, given samples from a known distribution, samples from the desired distribution can be obtained by applying a series of sufficiently complex invertible transformations to samples from known distribution.

The distribution of generated audios were obtained from sample audios conditioned on the mel-spectrogram in affine coupling layers. *WN* in Figure 2. contains layers of dilated convolutions with gated tanh nonlinearities, as well as residual connections and skip connections as a transformation of mel spectrogram for conditioning.

The 1x1 invertible convolutional layer before each affine coupling layer was implemented by initializing the weights to be orthonormal and hence invertible. This operations of passing through invertible counvolutional layer and then affine coupling layer was called "flow", which was concatenated multiple times to form the model structure.

2.3.2 MelGAN

MelGAN [7] is a non-autoregressive model built for text-to-speech task. MelGAN was also used for some similar music synthesis research, such as musical timbre transfer on drum tracks [8] and singing voice generation [4]. The model is extremely fast compared to similar models like WaveGlow. It models raw audio waveforms conditional on mel-spectrogram which is similar to WaveGlow. MelGAN is the first work that successfully trained GANs to convert spectrogram into raw audio without additional distillation or perceptual loss functions and still yielded high quality synthesized audio.

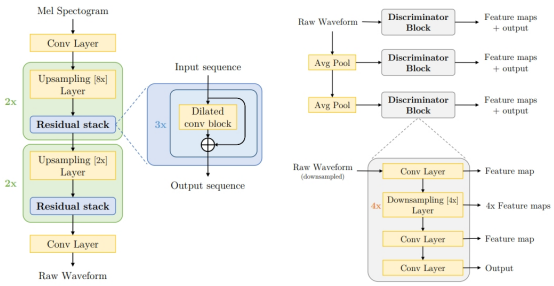


Figure 3. MelGAN

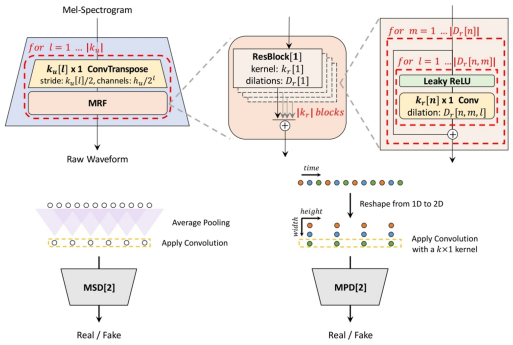


Figure 4. HiFi-GAN

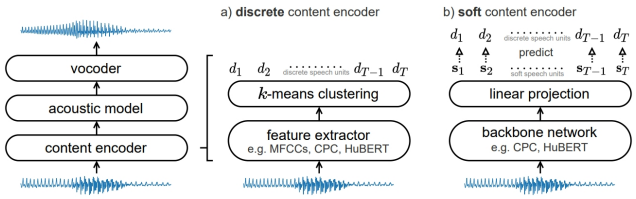


Figure 5. Hifi-GAN-VC

For generator, each transposed convolutional layer is followed by a stack of residual blocks with dilated convolutions. This model structure makes temporally far output activations of each subsequent layer has significant overlapping inputs to put an inductive bias that there is long range correlation among the audio timesteps. Receptive field of a stack of dilated convolution layers increases exponentially with the number of layers which increase the ability of generator for producing longer sequence.

The discriminators were designed as 3 multi-scaled architecture. They shared the same structure, but one operated on the raw audio and the other two operated on raw audio downsampled by a factor of 2 and 4 respectively. This design makes model learn different features for different frequency range of the audio which is appropriate for our multi-instruments sources.

2.3.3 Hifi-GAN

HiFi-GAN handled end-to-end speech synthesis not only by effectively modeling the long-term correlation of the speech waveform, but also by effectively modeling the periodic mode of the speech waveform. Besides, it achieves real-time and high-fidelity speech waveform generation. Moreover, as one of the most advanced vocoder networks, HiFi-GAN was used as the backend by many speech synthesis systems to restore the predicted mel spectrgram of speech.

The generator in Hifi-GAN is a fully connected convolutional network composed of transposed convolutional layers and multi-receptive field fusion module, which observes patterns of various lengths in parallel. For discriminator, in addition to multi-scaled subdiscriminator from MelGAN, there's another multi-period subdiscriminator to capture different implicit structures by looking at different parts.

2.3.4 Hifi-GAN-VC

Kamper et al. [26] proposed soft speech units for voice conversion and describe a method to learn them from discrete units. There are three parts in the voice conversion system. The first part is discrete and soft content encoder which were trained to transform audio into discrete and soft units with different feature extractor, such as hidden-unit BERT (HuBERT) and contrastive predictive coding (CPC). Then these units would be transformed into mel-spectrogram by acoustic model and passed through vocoders. In their study, HiFi-GAN was used as vocoder models.

2.4. Experiments

In transfer learning, we used the provided pretrained weights for all models by their corresponding authors. To evaluate the generality of HiFi-GAN to the mel-spectrogram inversion of unseen sequences, the dataset was split into trainging and validation set in 10-fold cross validation. The synthesis speed was measured on environments with a single NVIDIA GeForce GTX 1080Ti GPU and Intel Xeon Processor E5-2630 v4 in UCSD Research Cluster. The networks were trained using the Adam optimizer with weight decay 0.01. The learning rate decay was scheduled by a 0.999 factor in every epoch with an initial learning rate of 2×10^{-4}.

3. Results

3.1. Generated mel spectrogram

Instead of extracting vocals from audios, alternatively, the idea in this project is to utilize GANs to generate audios

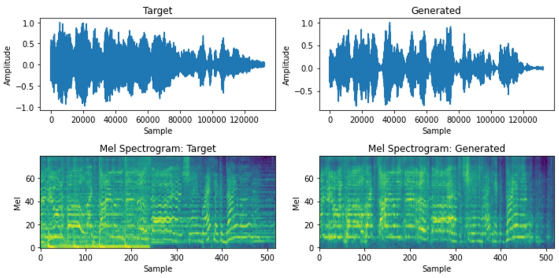


Figure 6. Result of waveforms and mel spectrogram by WaveGlow (left) target music audio (right) generated vocal

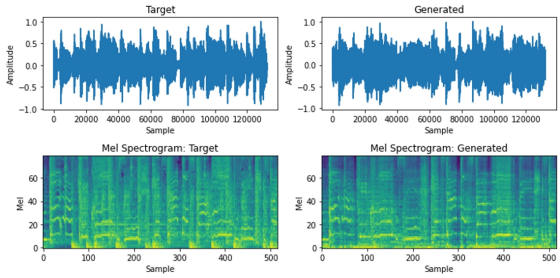


Figure 7. Result of waveforms and mel spectrogram by HiFi-GAN (left) target music audio (right) generated vocal

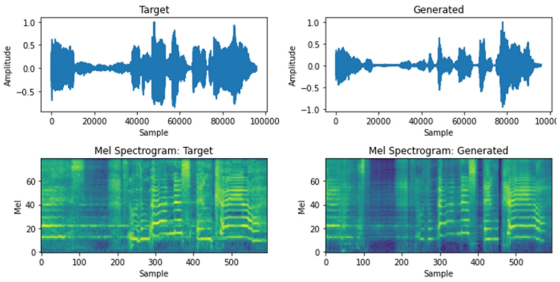


Figure 8. Result of waveforms and mel spectrogram by HiFi-GAN-VC (left) target music audio (right) generated vocal

containing vocals only. [1] Therefore, these models should have ability of differentiating vocals from multi-frequency ranges and non-vocal frames which contains no vocals.

As we can see from figures below, patterns of vocals and the corresponding harmonics in mel spectrogram were generally captured and preserved by all the generators. The energy of background instruments were attenuated to lower level compared to vocals. This is consistent with our expectation that generated audios would contained attenuated instruments(ideally, to very low level) and good perceptual quality of generated fake vocals. This validate our works with transfer learning from speech synthesis to MSS task.

As mentioned in Methods, using mel filter banks can condense information in spectrogram into low frequency range. Since human vocals were distributed in relatively lower frequency range as reported [1], the condensation by mel filter bank is effective to preserve human vocals and attenuate instruments like piano and guitar in high frequency range. This can also be observed in generated audios by all the GAN-based models we used in this project.

3.2. Audio quality

Unlike previous researches in audio synthesis or style transferring, audio quality was not evaluated in metrics like mean opinion score (MOS), like how MelGAN and Hifi-GAN were evaluated in their articles. The reason is that most common metrics for generative model in music or speech applications are generally **subjective** metrics [6, 7, 17, 28]. For example, MOS hearing test is to randomly assign generated sequences to raters for ratings on pleasantness on a five-point scale for measure. These metrics are difficult to be applied and measured in this project, so they were not used for evaluation.

The results of generated audio show that all the models outperform WaveGlow in the end-to-end setting as mentioned in [6]. We can observe that the generated audios by WaveGlow contain more noises compared to other models because the weak modeling power of the bipartite structure needs a larger number of transformation in order to estimate the distribution of complex music waveforms [13]. Also, its architecture is overly computational expensive for CPU/GPU-based inference with a constrained memory budget which is a drawback to build our proposed applications.

For Hifi-GAN and Hifi-GAN-VC, the sounds effects and noise level is much better than WaveGlow, and HiFi-GAN-VC is much better in most of tracks than HiFI-GAN. Most energy of instruments could be eliminated in low rhythmic tracks, such as country music or lyrical song. However, the generated audios were a little bit "twisted", which means the produced audios sound like recordings using low quality microphone. In addition, the performance of Hifi-GAN and Hifi-GAN-VC is weak for audios containing strong beats and fast rhythm, such as rap or rock music. Unlike speech application, the periodicity in pop music tracks is not strong because of variant rhythms of background instruments, so the ability of multi-period discriminator in HiFi-GAN would become a constraint for music audio synthesis. One potential solution might be manipulating hyperparameters in multi-period discriminator, such as increasing length of period to capture longer dependencies across sequences.

Finally, MelGAN gives the most promising results for our MSS task. With multi-scaled architecture in discriminator, the model can handle variant frequency range,

[1] Some generated samples in this project can be find here: https://github.com/wf1497c/-audio-samples-Mel2Seq-VC

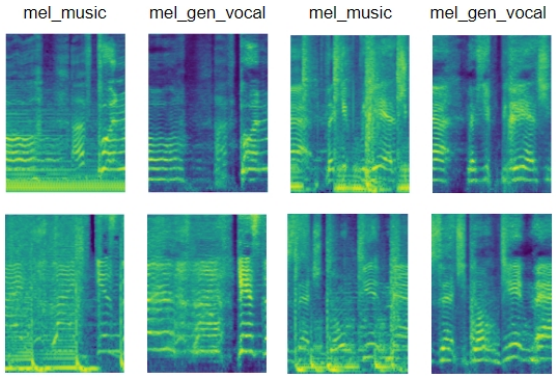


Figure 9. Result of mel spectrograms of audio and the corresponding generated mel spectrograms by MelGAN for different genres (upper-left) Lyrical song (upper-right) Nu-Metal rock (lower-left) Arena rock (lower-right) Modal jazz

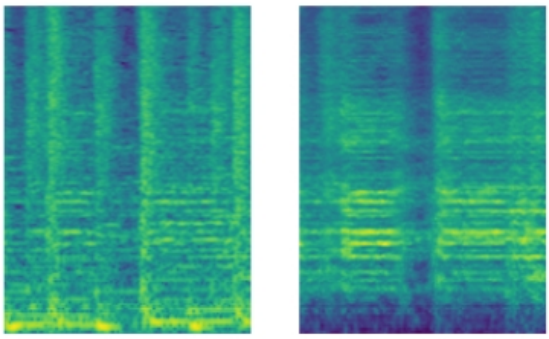


Figure 10. Mel spectrograms in non-vocal frame in MelGAN (left) music audio (right) generated vocal

strong beats, and different pitch/rhythm scenario. Also, it contained siamese networks, which are helpful to avoid checkerboard artifacts [12], and it worked well especially for transferring audio style. Another observation is its ability of handling different genres and styles of music audios, which is not found for other models. As a result, MelGAN will be used to develop our application for vocal separation and related parameters will be discussed in future study.

In our experiments, one issue in MelGAN is "non-vocal" frames, which is time frames only containing background instruments and no vocals. If instruments in these frames contained strong energy, MelGAN would sometimes transformed these instruments sounds into human vocals, which is an unexpected outcome. This happened especially for genres like heavy metals and should be fixed in future study.

3.3. Limitation

Due to limitation of our hardware, only small batch sizes (8-16) could be fitted into GPU memory and this makes

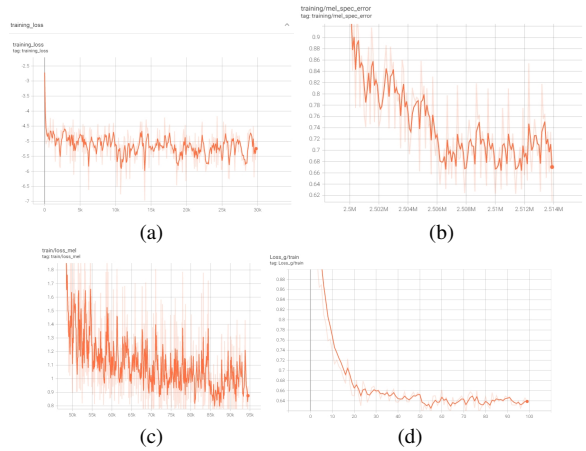


Figure 11. Training curves (a) WaveGlow (b) Hifi-GAN (c) Hifi-GAN-VC (d) MelGAN

fine-tuning tasks more challenging. The learning curves were shown in Figure. 11 and we can observe fluctuations of loss curves during training under adversarial training strategy which depends heavily on the abilities of generators and discriminators. Especially, large fluctuations in HiFi-GAN and HiFi-GAN-VC might be caused by the multi-period discriminators as mentioned above. The early stopping criterion was set depending on validation loss and final epochs are around 150.

Another limitation is the pre-trained models from authors. Since the parameters were set for their purposes, it's hard to modify the models' architectures or manipulate related hyperparameters for flexible usages. Intuitively, the length of input segments to discriminators and generators should be larger than the case of speech because patterns of waveforms in vocals and instruments were continuous for longer time intervals. However, these ideas were not validated and discussed in this project.

4. Conclusion

In this project, we have validated the potential values of GAN-mel spectrogram methods in MSS task. To best of our knowledge, this project is the first research using MelGAN and Hifi-GAN as vocal separator for pop music source separation. The usage of mel filter bank makes model learn patterns of vocals in low dimensional feature spaces and enhance distinguishability of components in our music tracks. However, problems of non-vocal frames and strong beats cases were also be reported in this project and more study to handle these is necessary to build real applications. In future study, MelGAN and its derived models will be developed for vocal separator of pop music tracks.

References

[1] Tech stuff - frequency ranges. https://www.zytrax.com/tech/audio/audio.html. 1, 5

[2] Dittmar C. Abeßer J. Müller M. Balke, S. Data-driven solo voice enhancement for jazz music retrieval. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 196-200). IEEE. 1

[3] Miron M. Janer J. Gómez E. Chandna, P. Monoaural audio source separation using deep convolutional neural networks, 2017. In International conference on latent variable analysis and signal separation (pp. 258-266). Springer, Cham. 2

[4] Cui C. Chen F. Ren Y. Liu J. Zhao Z. ... Wang Z. Huang, R. Singgan: Generative adversarial network for high-fidelity singing voice generation. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 2525-2535). 3

[5] Nieto O. Jin Z. Kandpal, N. Music enhancement via image translation and vocoding. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3124-3128). IEEE. 1, 2

[6] Jaehyeon Kim Kong, Jungil and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems 33 (2020): 17022-17033. 2, 3, 5

[7] Kumar R. de Boissiere T. Gestin L. Teoh W. Z. Sotelo J. ... Courville A. C. Kumar, K. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019. Advances in neural information processing systems, 32. 2, 3, 5

[8] Keon Ju. Lee. Computer evaluation of musical timbre transfer on drum tracks. Diss. 2021. 3

[9] Chen J. Hou H. Li M. Li, T. Sams-net: A sliced attention-based neural network for music source separation, 2021. In 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 1-5). IEEE. 1

[10] Chen Z. Yoshioka T. Luo, Y. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 46-50). IEEE. 2

[11] Colin Raffel Dawen Liang Daniel PW Ellis Matt McVicar Eric Battenberg McFee, Brian and Oriol Nieto. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, pp. 18-25. 2015. 2

[12] V. Dumoulin O., Augustus and C. Olah. Deconvolution and checkerboard artifacts. Distill 1.10 (2016): e3. 6

[13] Lim H. Byun K. Hwang M. J. Song E. Kang H. G. Oh, S. Excitglow: Improving a waveglow-based neural vocoder with linear prediction analysis. 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020. 5

[14] Kim T. Lee K. Kwak N. Park, S. Music source separation using stacked hourglass networks, 2018. arXiv preprint arXiv:1805.08559. 2

[15] Peter Balazs Perraudin, Nathanaël and Peter L. Søndergaard. A fast griffin-lim algorithm. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2

[16] Oh E. Choo K. Sung H. Jeong J. Osipov K. Francois H. Porov, A. Music enhancement by a novel cnn architecture, 2018. In Audio Engineering Society Convention 145. Audio Engineering Society. 1

[17] Rafael Valle Prenger, Ryan and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2, 3, 5

[18] Liutkus A. Stöter F. R. Mimilakis S. I. Bittner R. Rafii, Z. Musdb18-a corpus for music separation, 2017. 1

[19] C. Ravanelli M. Cornell S. Bronzi M. Zhong J. Subakan. Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 21-25). I. 1, 2

[20] Das R. K. Li H. Sharma, B. On the importance of audio-source separation for singer identification in polyphonic music, 2019. In INTERSPEECH (pp. 2020-2024). 1

[21] Pang R. Weiss R. J. Schuster M. Jaitly N. Yang Z. ... Wu Y. Shen, J. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. IEEE international conference on acoustics, speech and signal processing (ICASSP). 2

[22] Ewert S. Dixon S. Stoller, D. Wave-u-net: A multi-scale neural network for end-to-end audio source separation, 2018. arXiv preprint arXiv:1806.03185. 2

[23] Ravanelli M. Cornell S. Bronzi M. Zhong J. Subakan, C. Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25). I. 2

[24] Goswami N. Mitsufuji Y. Takahashi, N. Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In 2018 16th International workshop on acoustic signal enhancement (IWAENC) (pp. 106-110). IEEE. 2

[25] Porcu M. Giron F. Enenkl M. Kemp T. Takahashi N. Mitsufuji Y. Uhlich, S. Improving music source separation based on deep neural networks through data augmentation and network blending. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 261-265). IEEE. 1

[26] Carbonneau M. A. Zaïdi J. Baas M. Seuté H. Kamper H. van Niekerk, B. A comparison of discrete and soft speech units for improved voice conversion. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022. 2, 3, 4

[27] Li H. Toda T. Vijayan, K. Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes, 2018. IEEE Signal Processing Magazine, 36(1), 95-102. 1

[28] Eunwoo Song Yamamoto, Ryuichi and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. 2, 5

[29] Yu Y. Gao Y. Chen X. Li W. Zhang, X. Research on singing voice detection based on a long-term recurrent convolutional network with vocal separation and temporal smoothing, 2020. Electronics, 9(9), 1458. 1