2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI) | 978-1-6654-1270-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/BDAI52447.2021.9515219

Super-Resolution for Music Signals Using Generative Adversarial Networks

Jinhui Dai, Yue Zhang, Pengcheng Xie, Xinzhou Xu*
School of Internet of Things
Nanjing University of Posts and Telecommunications
Nanjing, P. R. China
Email: xinzhou.xu@njupt.edu.cn

***Abstract*—Super-Resolution (SR) refers to increasing the resolution of a signal in a variety of ways, conventionally employed in the field of image enhancement. Compared with the endeavors for super-resolution in image processing, music signals require supper-resolution to improve their quality or adapt to communication in narrow-band channel, which is also regarded as bandwidth expansion. To this end, we shed light on super-resolution for music signals using the deep learning strategy of Generative Adversarial Networks (GANs). The proposed approach feeds Shot-Time Fourier Transform (STFT) features of low-band signals to the GAN, expecting to obtain their high-band information through jointly considering content and adversarial losses. Then, we carry out experiments on MUSDB18 dataset using mixtures of music sources, in order to show the performance of the proposed approach. The experimental results indicate that the proposed approach achieves better super-resolution performances compared with interpolation and some conventional deep-learning strategies.**

***Keywords*—super-resolution, generative adversarial networks, music processing, bandwidth extension**

I. Introduction

Super-Resolution (SR) usually appears in image processing to achieve clearer or denser pixels within static images or video signals [1-3]. This can be categorized as a direction in audio enhancement [4], which mainly focuses on removing background noise and performing further bandwidth expansion [5]. On the research of bandwidth expansion for super-resolution, it is possible to generate high-frequency details using low-band information, in order to improve the quality of audio signals. For the application on super-resolution in speech signals, the expanded high frequency can improve the clarity of pronunciation [6]. Typically, it is also applicable to perform super-resolution in music signals, considering the requirements of auditory perception for high-fidelity music.

The existing research of audio super-resolution includes two aspects, namely shallow and deep pipelines. The shallow pipeline employs classical machine-learning models, including Gaussian Mixture Model (GMM) [7] and Hidden Markov Model (HMM) [8]. The deep pipeline typically refers to the inclusion of Deep Neural Networks (DNNs) to learn complex models for bandwidth expansion[9]. Then, *Kuleshov et al.* proposed to use Convolutional Neural Networks (CNNs) for audio super-resolution[10]. Afterwards, *Eskimez et al.* introduced a Generative Adversarial Network (GAN)-based strategy to further improve the performance of super-resolution for speech signals[11,12]. However, despite of these works on audio super-resolution, it is still difficult for music signals to perform super-resolution, due to their properties on mixed sources and flexibility.

Therefore in this work, we investigate super-resolution for music signals using the deep generative model of GAN. The proposed system utilizes Short-Time Fourier Transform (STFT) features as the input and a GAN with deep structures for learning on its generator and discriminator. Within the training procedures, the network jointly considers content and adversarial losses to model the mapping from the low-band to high-band information. Then, we carry out experiments on the MUSDB18 dataset, in order to examine the performance of the proposed approach. The contributions of this paper can be summarized as: We propose a GAN-based approach for super-resolution on music signals using the log-power spectrogram of STFT features, and perform corresponding experiments on real-world data.

The remainder of this paper is organized as follows. We introduce the work related to the research in Section II, the experimental method was described in Section III, and the experimental data and analysis were presented in Section IV, and finally the paper is summarized in Section V.

II. Related Works

A. Music Enhancement

Conventional music-enhancement approaches usually coordinate two topics: Separating music signals from other signals[13] (i.e., noise interference), and improving the quality of a signal, so that to make it close to the music recorded in a studio. Existing research includes the technology of speech separation [14] and speech enhancement, as well as the separation of music to extract the instrumental components or the removal of human voices, to produce pure-music components of a song. The applications of music enhancement include enhanced mobile device recording, hearing aids, and teleconference systems. For a specific task in music enhancement, *Roblek et al.*[15] found that it is challenging to quantitatively compare the output perception quality of different methods or models, and proposed an automatic

*Corresponding author. This work was supported by the Natural Science Foundation of China under Grants No. 61801241, No. 61801236, and No. 61802206, the Natural Science Foundation of Jiangsu under Grant BK20180746, and the NUPTSF under Grants XZD2020111 and XYB2020270.

978-1-6654-1270-4/21/$31.00 ©2021 IEEE

metric for music enhancement based on the Fréchet Audio Distance (FAD) metric.

B. Deep Learning Based Bandwidth Expansion

Early research focused on the mapping of narrow-band signals to wide-band signals[16]. *Nakatoh et al.* proposed to use linear mapping for bandwidth expansion[17]. *Chennoukh et al.* proposed to use GMM[18,19], while *Bauer et al.* proposed to use HMM to consider [20-22]. All of these methods have achieved general results.

Prior to the appearance of deep learning, neural network had already made some achievements in the field of bandwidth expansion [23,24]. In the recent decade, deep-learning methods improve super-resolution performance for audio signals using diversity of network structures. Among the related works, the research on Convolutional Neural Networks (CNNs) has shown the effectiveness for image super-resolution. In this regard, *Kuleshev et al.* employed CNN to predict the high frequency part of audio[10]. Compared with regression models like DNN and CNN, the generative-network model of GAN introduces discriminator and generator in adversarial training for super-resolution on audio data, which was proved with better performance[11].

III. Methodology

Fig. 1 shows the complete process of audio super resolution processing. First, the signal is transformed into the log-power spectrogram and the phase spectrum through STFT. Second, the low-frequency and high-frequency parts are separated. In this procedure, the low-frequency log-power spectrogram is sent into the GAN network to obtain the high frequency parts, and the low frequency phase spectrum is flipped to obtain the high frequency parts. Finally, Inverse Short-Time Fourier Transform (ISTFT) is used to reconstruct the signal[11].

A. The Spectrogram Features

First, the data requires pre-processing for generating features for deep networks. We down-sample an arbitrary audio signal to 16kHz, and then apply STFT to one of its frames $x(t)$. Thus, the principle of STFT for $x(t)$ is represented to obtain its frequency-domain form as

$$X(f) = STFT(x(t)) = \int_{-\infty}^{+\infty} x(u)g(u-t)e^{-j2\pi f}du \ , \qquad (1)$$

where $x(u)$ is the source signal and $g(u-t)$ is the window function. We choose the Hamming window as the window function, with its window length as 32ms (512 sampling points) and the window shift as 16ms (256 sampling points).

Then, we extract the log-power spectrogram $ln\,|\,X\,|^2$ and phase spectrogram $\angle X$ for each frame in the function $STFT(X)$. In view that the audio X is conjugate symmetric, we choose the first 257 spectral points of each frame in the log-power spectrogram as the effective data. Furthermore, we take the first 129 spectral points as the Low-Band (LB) features, whilst the other 128 points as the High-Band (HB) features of the total 257 points. For the phase $\angle X$ of the spectrogram, it is similar that the first 129 spectral points are selected as the LB features. The other spectral points are obtained by flipping the first 129 spectral points directly. Fig. 1 shows the diagrammatic overview of the proposed approach, including the steps of STFT, phase and log-power spectrogram processing, and ISTFT.

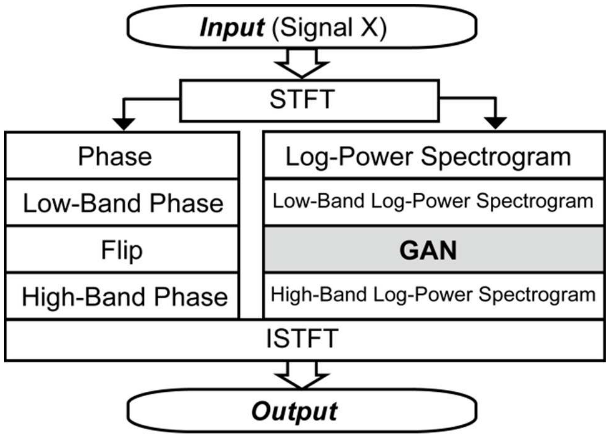


Fig. 1. The digrammatic overview of the proposed system

B. The Network Architecture

Then, we focus on the GAN procedure with its network architecture presented in Fig. 2, containing the generator and the discriminator. Within this adversarial structure, the generator module aims to generate HB information from LB signals, while the discriminator module is designed for adversarial processing on the HB signals. We set the time-step to 32, which can be seen as the usage of 32 consecutive frames in the time domain.

The generator includes the convolutional layers with the shortcut connections of Residual Network (ResNet), for three pairs of layers[11,25]. The discriminator contains convolutional and dense layers to separate predicted and true HB signals. Note that all the convolutional layers in the generator and the discriminator utilize the one-dimensional convolution (Conv1D) strategy[11], due to its conventional usage in audio signal processing and its lower computational complexity, compared with the two-dimensional convolutional-layer cases. Note that the kernel size for the eight Conv1D layers in the generator are set to 7, 5, 3, 3, 3, 5, 7, and 9 respectively, while for the discriminator the sizes are 7, 5, and 3 respectively.

It is also worth noticing that each convolutional layer of the generator is followed by a Batch Normalization (BN) layer and a Leaky Rectified Linear Unit (LeakyReLU) layer[11], where the slope of the LeakyReLU layer is set to 0.2. Each convolutional layer in the discriminator is followed by a LeakyReLU layer, also with the slope of 0.2.

C. The Loss Function

The loss function of the proposed super-resolution approach can be defined as

$$L^{SR} = aL^{content} + bL^{adversarial} \ , \qquad (2)$$

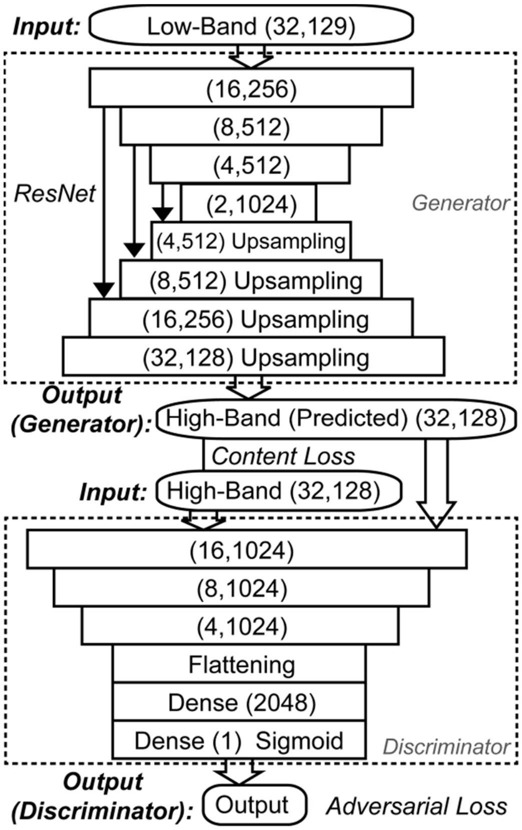


Fig. 2. The generative adversarial network architecture used in the propsoed approach, including generator and discrimintator modules

where $a,b \geq 0$ are the weighting coefficients and the content loss $L^{content}$ for the generator is defined as

$$L^{content} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \ , \qquad (3)$$

where y_i represents the original HB signal, while $\hat{y}_i$ is the output signal of the generator, and N is the total number of training samples. The adversarial loss $L^{adversarial}$ for the discriminator is set as

$$L^{adversarial} = log(1 - D(G(x))) \ , \qquad (4)$$

where $G(\bullet)$ represents the mapping of the generator, while $D(\bullet)$ is the discriminator mapping, with x representing the input LB signal.

IV. Experiments

A. Experimental Setups

We employ the MUSDB18 dataset in the experiments[26,27], containing mixed sources of music pieces with the sampling rate of 16kHz. The dataset consists of 150 full-length music tracks, with sub-track audio sources of drums, bass, vocals, and other audio tracks from different genres. The five sound tracks for each music piece in the dataset include one mixed sound track, three instrumental sound tracks, and one voiced sound track. We choose all the 150 music pieces in the dataset with mixed tracks in the experiments, and divide them into music-piece-independent training, validation, and test sets with a 6:4:5 ratio. This leads to 886 012, 569 284, and 778 736 frames for the three sets respectively.

We employ the Adaptive moment estimation (Adam) optimizer in training the network[11,12], including the polynomial learning rate decay scheduling with rate of 0.5. The learning rate of the discriminator is fixed at 10^{-7}, while the initial learning rate of the generator is 10^{-4}. In order to solve the problem that the generator loss rapidly increases and the discriminator loss quickly returns to zero in the early stage, we set a in Eq.(2) fixed at 0.1 and the initial value of b at 0.001. In the training procedure, we train the dataset for 30 epochs and decrease the learning rate of the generator and increase b in Eq.(2) in each epoch. In addition, the batch size of the network is set to 64. .

B. Evaluation Metrics

We choose Log-Spectral Distance (LSD) and Segmental Signal-Noise Ratio (SegSNR) as the evaluation metrics in the experiments[10-12]. LSD describes the distance between the results of two approaches, while SegSNR indicates the ratio between signal and noise, defined as

$$L_{LSD} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(X\left(l,k\right) - \hat{X}\left(l,k\right)\right)^2} \ , \qquad (5)$$

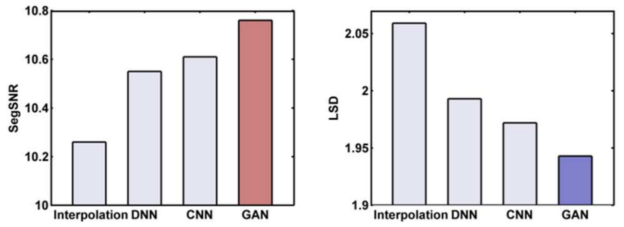
$$SegSNR = \frac{1}{L}\sum_{l=1}^{L} 10\lg\frac{\sum_{k=1}^{K}\left(X(l,k)\right)^2}{\sum_{k=1}^{K}\left(X(l,k) - \hat{X}(l,k)\right)^2} , \qquad (6)$$

where L and K represent the numbers of the frames and the frequency points within a sample, respectively. $X\left(l,k\right)$ corresponds to the original HB signal without and with STFT processing respectively, while $\hat{X}(l,k)$ corresponds to the predicted HB information.

C. Experimental Results and Analysis

We perform experiments on the case of mixed music sources and compare the experimental results with three control methods: linear interpolation, DNN, and CNN. The DNN contains four hidden layers with 256 nodes for each hidden layer. The CNN approach employs the three Conv1D layers including 256, 512, and 1 024 convolutional filters for these layers with the size of 7, 5, and 3 respectively. Note that both of the DNN and CNN are trained on frame-level samples, due to the better performance compared with the multi-frame setting.

First, we focus on two-time Super-Resolution (SR×2) to reconstruct original music signals. In view of the SegSNR and LSD results presented in Fig. 3, the GAN achieves the highest SegSNR and the lowest LSD among the approaches in the comparison. This indicates that it is possible to result in better performance with the help of generative architectures for super-resolution in music signal processing.



(a) (b)

Fig. 3. The experimental results for the approaches of interpolation, DNN, CNN, and GAN using the (a) SegSNR and (b) LSD metrics.

Then, we add the experimental results on four-time Super-Resolution (SR×4) to further investigate the super-resolution performance with the LB information of 65 frequency points, as shown in Table I and II with the metrics of SegSNR and LSD respectively. It is learnt from the tables that the three deep-learning methods perform better than the interpolation approach for both of the SR×2 and SR×4 tasks. Within the three deep methods, GAN achieves the best performance with its SegSNR as 10.76 and 9.11 for SR×2 and SR×4 tasks respectively, while the results are 1.943 and 2.130 when using LSD.

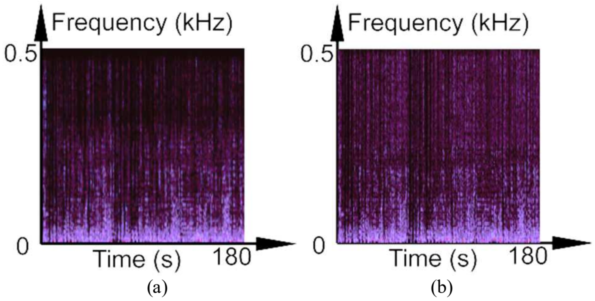
TABLE I. SUPER-RESOLUTION RESULTS USING THE SEGSNR METRIC

Approaches\Resolution	SR×2	SR×4
Interpolation	10.26	8.07
DNN	10.55	8.94
CNN	10.61	9.07
GAN	**10.76**	**9.11**

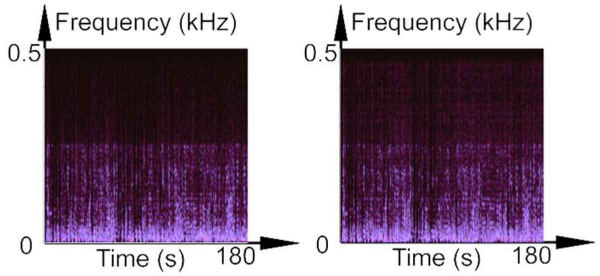
TABLE II. SUPER-RESOLUTION RESULTS USING THE LSD METRIC

Approaches\Resolution	SR×2	SR×4
Interpolation	2.059	2.406
DNN	1.993	2.166
CNN	1.972	2.143
GAN	**1.943**	**2.130**

Finally, we examine the wide-band spectrograms of the original signal and the signals using DNN, CNN, and GAN approaches respectively for the SR×2 task, as presented in Fig. 4 and Fig. 5, with respect to two examples with the time-domain length of 180s. It is seen from the figures that the proposed approach may do better in generating high-frequency information compare with the other approaches.

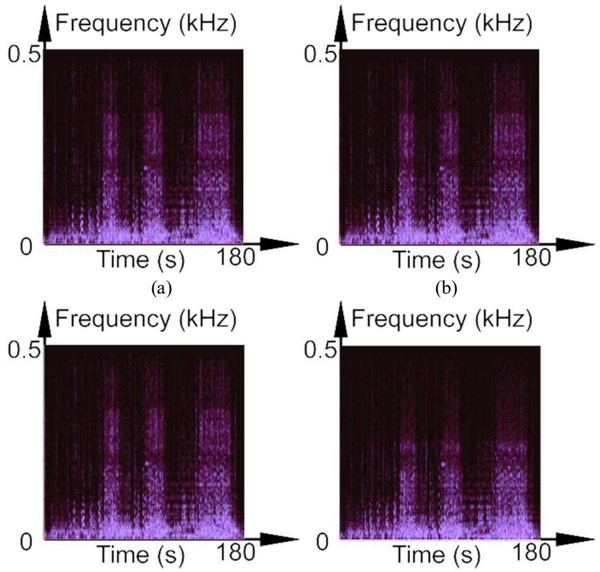


(a) (b)



(c) (d)

Fig. 4. Wide-band STFT-spectrograms for Example 1 for (a) the original signal, (b) GAN, (c) DNN, and (d) CNN



(a) (b)

(c) (d)

Fig. 5. Wide-band STFT-spectrograms for Example 2 for (a) the original signal, (b) GAN, (c) DNN, and (d) CNN

V. CONCLUSIONS

In this paper, we investigated super-resolution for music signals using a Generative Adversarial Network (GAN) approach in order to perform bandwidth expansion in producing high-fidelity music. In this regards, we performed experiments on MUSDB18 dataset with mixed sources of multiple sound tracks. The experimental results show that the proposed GAN is superior to the conventional approaches including DNN and CNN. Our future works may contain two aspects as follows: First, we aim to further improve the performance on the basis of the proposed approach through considering attention tricks. In addition, it is also possible to investigate super-resolution for music signals in cross-domain conditions.

REFERENCES

[1] C. Ledig, L. Theis, and F. Huszár, "Photo-realistic single image super-resolution using a generative adversarial network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681-4690, 2017.

[2] X. Zhu, Z. Li, J. Lou, and Q. Shen, “Video super-resolution based on a spatio-temporal matching network,” Pattern Recognition, vol. 110, pp. 107619, 2021.

[3] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, “Deep video super-resolution using HR optical flow estimation,” IEEE Transactions on Image Processing. vol. 29, pp. 4323-4336, 2020.

[4] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 55-66, 2015.

[5] P. Smaragdis and B. Raj, “Example-driven bandwidth expansion,” 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 135-138, 2007.

[6] D. Sinha, S. AJ Ferreira, and EV Harinarayanan, “A Novel Integrated Audio Bandwidth Extension Toolkit (ABET),” Audio Engineering Society, 2006.

[7] K. Park and H. S. Kim, “Narrowband to wideband conversion of speech using gmm based transformation,” 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol.3, pp. 1843-1846, 2000.

[8] P. Jax and P. Vary, “Artificial bandwidth extension of speech signals using mmse estimation based on a hidden Markov model,” 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP), vol. 1, pp. 1-1, 2003.

[9] K. Li and C. H. Lee, “A deep neural network approach to speech bandwidth expansion,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 4395–4399.

[10] V. Kuleshov, S. Enam, and S. Ermon, “Audio super resolution using neural networks,” in ArXiv, 2017.

[11] S. E. Eskimez and K. Koishida, “Speech super resolution generative adversarial network,” ICASSP 2019-2019 IEEE Inter- national Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3717–3721, 2019.

[12] S. E. Eskimez, K. Koishida, and Z. Duan, “Adversarial training for speech super-resolution,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 347–358, 2019.

[13] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, “Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs,” IEEE Transactions on Audio Speech & Language Processing, vol. 15, no. 5, pp. 1564-1578, 2007.

[14] S. Makino, H. Sawada and T. W. Lee, “Blind Speech Separation,” Springer Netherlands, 2007.

[15] D. Roblek, K. Kilgour, M. Sharifi and M. Zuluaga, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” In Proceedings of INTERSPEECH 2019, pp. 2350-2354, 2019.

[16] B. Iser and G. Schmidt, “Bandwidth extension of telephony speech,” Speech and Audio Processing in Adverse Environments, pp. 135–184, 2008.

[17] Y. Nakatoh, M. Tsushima, and T. Norimatsu, “Generation of broadband speech from narrow-band speech using piecewise linear mapping,” in Fifth European Conference on Speech Communication and Technology, 1997.

[18] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, “Speech enhancement via frequency bandwidth extension using line spectral frequencies,” IEEE International Conference on Acoustics, Speech, & Signal Processing 2001, pp. 665–668, 2001.

[19] H. Seo, H. Kang, and F. Soong, “A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 6087-6091, 2014.

[20] G. Chen and V. Parsa, “HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies,” 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, vol. 1, pp. 1-709, 2004.

[21] P. Bauer and T. Fingscheidt, “An HMM-based artificial bandwidth extension evaluated by cross-language training and test,” 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, pp. 4589-4592, 2008.

[22] G. Song and P. Martynovich, “A study of hmm-based band- width extension of speech signals,” Signal Processing, vol. 89, no. 10, pp. 2036–2044, 2009.

[23] B. Iser and G. Schmidt, “Neural networks versus codebooks in an application for bandwidth extension of speech signals,” in Eighth European Conference on Speech Communication and Technol- ogy, 2003.

[24] J. Kontio, L. Laaksonen, and P. Alku, “Neural network-based artificial bandwidth expansion of speech,” IEEE Transactions on Audio Speech and Language Processing, vol. 15, no. 3, pp. 873–881, 2007.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 630-645.

[26] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, and R. Bittner, “MUSDB18-a corpus for music separation,” December 2017.

[27] F. Stöter, A. Liutkus, N. Ito, “The 2018 signal separation evaluation campaign,” International Conference on Latent Variable Analysis and Signal Separation. Springer, 2018, pp. 293-305