Multi-band Masking for Waveform-based Singing Voice Separation

Panagiotis Papantonakis[1], Christos Garoufis[1,2], and Petros Maragos[1,2]
[1]*School of ECE, National Technical University of Athens*, 15773 Athens, Greece
[2]*Robot Perception and Interaction Unit, Athena Research Center*, 15125 Maroussi, Greece
panpapantonakis@gmail.com, cgaroufis@mail.ntua.gr, maragos@cs.ntua.gr

***Abstract*—Singing Voice Separation (SVS) is the task of isolating the vocal component from a given musical mixture. In recent years there has been an increase in both the quantity and quality of SVS techniques that operate in the waveform domain and have an encoder-separator-decoder structure, where the separator processes a latent representation of the waveform produced by the encoder. In this work, we propose a parallel multi-band modification for this family of architectures, that splits the latent representation provided by the encoder in multiple sub-bands and then processes each band in isolation, using multiple separators, so as to better exploit the internal correlations of each sub-band. We investigate the effect of our proposed modification on Conv-TasNet, a widely used architecture adhering to the encoder-separator-decoder paradigm. The results indicate that the proposed modification improves the overall performance without altering the network size, and offer insights on its scaling capabilities as well as its applicability in other architectures that follow this general paradigm.**

***Index Terms*—singing voice separation, Conv-TasNet, parallel separators, music source separation**

I. Introduction

Source separation is the task of decomposing a given mixture signal into the source signals that constitute it. In the context of audio processing, singing voice separation is the task of separating the vocal track and the instrumental accompaniment from a musical mixture. With the development of deep learning, and the availability of more data [1], fully supervised methods based on Deep Neural Networks (DNNs) have dominated the field [2], [3]. These methods can be split in two broad categories, based on whether they process the audio signal in its original form (waveform domain), or using a time-frequency (T-F or spectrogram) representation of the signal, such as the Short-Time Fourier Transform (STFT) magnitude.

The intuition behind the usage of the magnitude STFT for this task has led to a number of very successful network architectures [4]–[9]. However, the STFT still constitutes a generic signal transformation, not necessarily optimised for the task of source separation. Moreover, STFT-magnitude based methods omit the phase of the signal from the estimation of the sources. The most commonly used shortcut involves reconstructing the source signals using the phase of the mixture signal, or an approximation provided by the Griffin-Lim algorithm [10], thus discarding part of the signal information and hindering the overall separation potential.

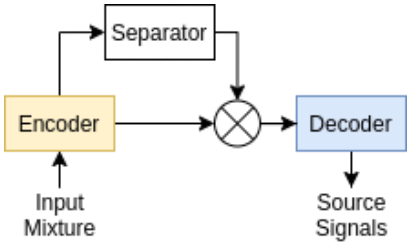

Fig. 1. The encoder-separator-decoder architecture.

On the other hand, waveform domain-based methods [11]–[16], bypass these issues, as they typically learn a transformation suitable for the separation task and utilise all the information provided by the signal. A considerable amount of these waveform-based methods follow an encoder-separator-decoder structure [12], [17]–[19] and separate the source signals by transforming the input mixture into a two-dimensional latent space, generating suitable masks for each source, applying them on the learned latent space and transforming the results back to the waveform domain. These architectures have been shown experimentally to achieve a higher upper bound of performance, compared to their STFT-based counterparts [20].

Multi-band audio processing has proven to be successful in a number of tasks, such as generative audio synthesis [21], [22] and STFT-based audio source separation [8], [9]. In this work[1] we propose a multi-band extension to waveform-based architectures for audio source separation, in which the learned latent space is automatically split into multiple, uniquely separable bands, and each band is processed individually, using multiple separators in parallel. This work deviates from a number of publications in the literature [14], [19], [23]–[25] which either focus on utilizing an improved encoder structure, or fine-tuning the building blocks of the separator for time-domain encoder-separator-decoder architectures. We applied our proposed method in Conv-TasNet [12], a recently introduced architecture that follows the encoder-separator-decoder paradigm, and our results in the task of singing voice separation on the widely used MUSDB18 dataset [1] indicate that the proposed band-splitting paradigm could be successfully applied to a number of architectures that are based on the aforementioned learned latent space scheme.

This research was partially supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project Number: 7773).

[1]Github link: https://github.com/PanagiotisP/svs-multiband

The rest of the paper is organized as follows: In Sec. II, we present the original Conv-TasNet [12] architecture, while our proposed multi-band variant is introduced in Sec. III. The experimental framework we utilize to evaluate our proposed architectures is outlined in Sec. IV. We report on and discuss our findings in Sec. V, while Sec. VI contains our conclusions and proposes a number of future research directions.

II. REVIEW OF CONV-TASNET

The original Conv-TasNet [12] architecture consists of an encoder, a separation module and a decoder, as illustrated in Fig. 1. The encoder transforms overlapping segments of the mixture waveform $\mathbf{x} \in \mathbb{R}^{A \times T}$, where A is the number of audio channels (1 for mono, 2 for stereo) and T is the segment length in samples, into a latent, high-dimensional representation, $\mathbf{w} \in \mathbb{R}^{N \times T'}$, where T' is the length of the encoded representation in samples. As this transformation is learnable, it is possible to create representations of the signal that are more suitable for the separation process than other fixed transformations, such as the STFT. The encoder is implemented as an 1D convolutional layer, with an encoding dimension of $N = 256$, a relatively large kernel size of $L = 20$ samples, and a 50% stride of $L/2 = 10$ samples.

The separator processes the latent representation to generate, for each of the C sources in the mixture, a latent mask $\mathbf{M_c} \in \mathbb{R}^{N \times T'}, c = 1 \dots C$. These masks are applied multiplicatively to the encoded input, resulting in the encoded representations of the source signals. In [12], the separator is implemented as a multi-block, residual Temporal Convolutional Network (TCN) [26], that uses serially connected stacks of $S = 3$ sub-modules. In turn, these modules comprise of $D = 8$ depthwise separable convolutional blocks each, with an increasing dilation factor, $d_i = 2^{i-1}$, $i = 1, ..., D$, in order to capture multi-scale data patterns. Before and after the convolution operation, each convolutional block transforms the latent representation along the channel dimension between a bottleneck dimension, $B = 256$ and an internal channel dimension, $H = 512$.

Finally, the decoder transforms the masked latent representations of each source back to the waveform domain, resulting in the estimated segments of the source signals $\hat{s}_c$. This module uses a linear layer to change the number of channels from N to $A \cdot L$ and then uses an overlap-and-add method to restore the dimensionality of both the channel and the feature map dimensions of the source signals.

III. MULTI-BAND SEPARATION

In this section, we describe our proposed multi-band extension for Conv-TasNet. This was inspired by MMDenseLSTM [8], a time-frequency domain model that splits the mixture spectrogram into multiple frequency bands and processes each band individually, before combining their respective outputs. We attempt to create a similar structure, taking into consideration that the domain of operation differs (waveform instead of T-F) and the representation we use is derived via a learnable transformation, instead of the STFT, which is fixed, with known properties and interpretation.

A. Multi-Band Masking Separation Architecture

Figure 2 depicts the proposed model. In order to create multiple frequency bands, as in [8], we approached the latent representation as a T-F one. More specifically, regardless of the number of channels and the length of the input signal, the encoder produces a multi-channeled representation of 1D series of features. One can interpret this representation as having a "latent frequency" (channel) and a "latent time" (feature map) dimension, resulting in a "latent spectrogram". Of course, since the transformation that the encoder applies is learnable, it does not share the same interpretation as a deterministic transformation to the frequency domain.

Having made the above semantic remark we proceed in describing the proposed architecture. Our model starts off with the encoder, which is equivalent to the original [12]. The resulting latent representation $\mathbf{w} \in \mathbb{R}^{N \times T'}$ is split along the channel dimension to create Q uniquely exclusive feature map bands $\mathbf{w_i} \in \mathbb{R}^{N_i \times T'}, i = 1 \dots Q$, where N_i corresponds to the number of channels that are assigned to each band so that $N_i = N/Q$. Instead of a single separator, our model utilizes an ensemble of separators $\mathcal{S}_i$, $i = 1, ..., Q$, connected in parallel, each of which receives the feature map $\mathbf{w}_i$, and generates its respective sub-mask matrix $\mathbf{M_i} \in \mathbb{R}^{C \times N_i \times T'}$. By having the separators operate in smaller sub-space than the original, they can specialize in its properties and process the respective parts of the representation of the signal better.

These sub-mask matrices are then concatenated along the channel axis to restore the channel dimension of the encoded latent representation. After concatenation of the sub-mask matrices, the masks are applied multiplicatively to the latent representation of the mixture, to isolate each source signal in the latent space, and the decoder transforms the latent representations back to the waveform domain, providing us with the source signals. As in [12], the whole network is jointly trained, forcing thus the encoder to learn a latent space with sub-spaces exclusively tailored for each separator.

B. Full-band Masking Variant

In the proposed technique, each separator operates on a different sub-space than the others, as the assignment of channels to bands is mutually exclusive. In order to examine whether the separators work better in isolation, or they can benefit from "inter-band" information we propose an additional, "full-band", separator that processes the whole latent space. In this case, the total number of separators is $Q + 1$, with the last separator receiving all $N_{Q+1} = N$ channels of the latent representation and resulting in a $\mathbf{M} \in \mathbb{R}^{C \times N \times T'}$ mask. Since the total number of channels after concatenating the masks in the channel dimension is $2 \cdot N$, we incorporate a linear layer between the separators and the decoder to change the channel number back to N, so that we can multiply the resulting mask with the encoded mixture and retrieve the source signals.

IV. EXPERIMENTAL SETUP

A. Dataset

Following the majority of recent papers on music source separation, we evaluate our models on the MUSDB18 dataset

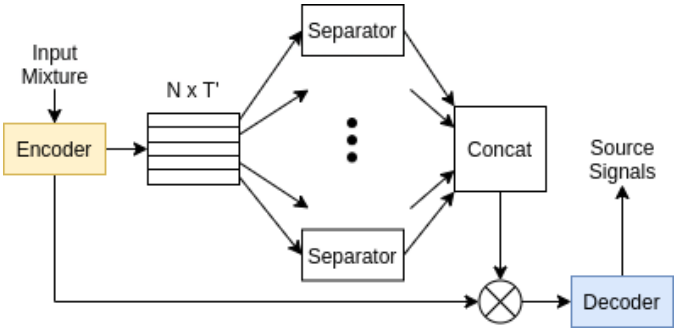

Fig. 2. Block diagram of our proposed multi-band separation modification. The N-channeled latent representation is split in the channel dimension in non overlapping bands, according to increasing channel index.

[1]. This dataset includes a total of 150 songs, in stereo format, at a sampling frequency of 44.1kHz, as well as the vocals, bass, drums and the rest of the accompaniment (denoted as "others") as individual tracks. Since we perform singing voice separation, we need a vocal component, which is directly provided by the dataset, and an accompaniment component, which is composed on the fly by adding the three instrumental streams. The dataset was split into training, validation and testing data according to the default configuration of 86 training, 14 validation and 50 test track songs.

Adapting the training to our resources, we use 2 second segments of dual-channel (stereo) signals, downsampled at 22.05kHz. As data augmentation, we apply channel and sign flip at random, as well as time-shift and shuffling of the sources between tracks of the same batch, similar to [27].

B. Training Configuration

We trained our models using the Adam optimizer, with a learning rate of 3e-4, for a maximum of 150 epochs and early stopping if 20 consecutive epochs pass with no decrease in the validation loss. The Mean Absolute Error (MAE) between the estimated and true source signals was used as the loss function. The batch size of each model was adapted according to its memory requirements, by always maximizing memory utilization of the used GPUs. Finally, the processing took place in 2 GeForce GTX 1080 Ti graphic cards.

C. Model Configurations and Variants

As our baseline model we use our reimplementation of the Conv-TasNet [12], with model hyperparameters set to the values mentioned in Sec. II. Also, for all model configurations involving multiple separators, the bottleneck dimension of each separator is equal to the number of channels it processes.

We call the baseline model M1. Regarding the multi-band models, we trained a model with $Q = 2$ bands (M2), a "full-band" model with $Q + 1 = 3$ bands (M3) and a model with $Q = 4$ bands (M4), to explore the scalability of the technique. Additionally, in order to investigate the ability of multiple separators to adapt to a predefined latent space, we train a model with $Q = 2$ bands, with the weights of the encoder and the decoder frozen at the respective values of the baseline, M1 (M5), as well as one with the same encoder and decoder as M5 (M6), but with the filters sorted in ascending order of base frequency, so that the two separator process solely the high- and low- frequency parts of the input, respectively.

Finally, in order to investigate whether the technique can co-operate with a more sophisticated front-end, that incorporates features from both waveform and time-frequency domains, we train two additional models that incorporate the stronger encoder/decoder described in [19]. In particular, the stronger encoder applies multiple 1-D convolutional layers with varying kernel sizes to capture features in multiple resolutions, and combines them with features derived from the magnitude STFT spectrogram of the signal, before processing the concatenated feature map with two 1x1 convolutional layers. This front-end can fully replace the original one, without the need of any adjustment to the separator. So, for this set of experiments we retrain the basic Conv-TasNet architecture from scratch, with its front-end replaced with the one described above (S1), and a second one (S2) that additionally incrorporates the two-band modification. The hyperparameters of the new modules are the same as in the original study [19].

D. Evaluation Protocol

In accordance to the protocol presented in [28], we evaluate our models against the Signal-to-Distortion-Ratio (SDR), Signal-to-Artifact-Ratio (SAR) and Signal-to-Interference-Ratio (SIR) metrics between estimated and true sources over 4-second segments. Regarding the computation of these metrics, it is assumed that the estimated source signal, $\widehat{s}_j$ is decomposed in 4 terms, corresponding to the true source signal, inter-source interferences, sensor noise and auditory artifacts, as: $\widehat{s}_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}$. The proposed decomposition is based on orthogonal projections of the source signals onto subspaces spanned by the source signals and/or the sensor noise. Therefore, the metrics are defined as:

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}$$

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}$$

$$\text{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}$$

SDR is considered to measure the overall quality of the separated signals, while SIR and SAR quantify the clarity of the separated sources and the existence of auditory artifacts in the sources, respectively. These metrics are calculated for both the estimated vocal and accompaniment components of each segment, using the museval python package which provides an implementation of the BSSEval metrics. The exact version used is BSSEval v4. In order to acquire a single value for each metric, we follow the median-of-medians protocol devised in [3]; first, the segment-wise scores are aggregated over each song by calculating their median, and then the median of the per-song scores is computed throughout the whole test set.

V. RESULTS AND DISCUSSION

Quantitative Model Comparison: Table I shows the evaluation results in terms of SDR, SIR and SAR. The two-band

TABLE I

MEDIAN SDR, SIR, SAR METRICS IN DB CALCULATED UPON 4-SECOND SEGMENTS IN MUSDB18 DATASET. HIGHER IS BETTER. BOLD DENOTES STATISTICALLY SIGNIFICANT IMPROVEMENT ($p = 0.01$) OVER THE RESPECTIVE BASELINE (M1 FOR M MODELS, S1 FOR S MODELS).

	M1	M2	M3	M4	M5	M6	S1	S2
SDR	5.81	**6.37**	5.94	6.05	**6.31**	**6.26**	6.39	6.36
Voc. SIR	14.13	14.25	14.23	14.61	**15.29**	**15.21**	14.39	14.92
SAR	6.59	**7.12**	6.78	6.98	**6.75**	**6.88**	6.82	7.09
SDR	11.78	**12.21**	11.76	11.66	**12.36**	11.91	12.23	12.03
Acc. SIR	16.01	**16.69**	16.01	16.04	**17.07**	**16.54**	17.57	17.51
SAR	14.24	**14.52**	14.37	14.10	14.11	14.29	14.20	14.07

TABLE II

DESCRIPTION AND NUMBER OF TRAINABLE PARAMETERS OF THE TRAINED MODELS.

Model	Description	#Params
M1	Baseline	6.6M
M2	2 Bands	6.58M
M3	2 Bands +1 Full-Band	12.97M
M4	4 Bands	6.71M
M5	2 Bands + Frozen enc/dec	6.56M
M6	2 Bands + Sorted enc/dec	6.56M
S1	Stronger enc/dec	7.32M
S2	Stronger enc/dec + 2 bands	7.31M

models M2 and M5 and achieve the best overall performance among M models, surpassing the metrics of the baseline, by a significant margin. This improvement over the baseline reaches 0.5 dB, in terms of vocal SDR. We note that this increase in performance is less than the one reported in [9], where application of the multi-band technique yielded a gain of approximately 1 dB in vocal SDR; however, no direct comparison can be made, since [9] was trained and evaluated on a subset of [1]. Regarding the other M models, the full-band model, M3, performs approximately equally to the baseline, the sorted frozen model, M6, performs better than the baseline but slightly worse than M2 and M5, while M4 performs similarly to the baseline regarding the accompaniment but scores better in terms of the vocal metrics, achieving a high SIR and SAR. On the other hand, the two models incorporating the encoder presented in [19] perform comparably to models M2 and M5, with no model achieving a clear edge in performance. Finally, we note that the results of baseline models M1 and S1 are mostly in agreement with those reported in [19], while using the normalized metrics devised in [29] yields similar results.

In order to assess the validity of our results, we performed the paired Wilcoxon signed rank test, over all metrics, between each model we developed and its respective baseline, at a statistical significance level of $p = 0.01$. We note that models M2, M5 and M6, which utilize 2 uniquely exclusive bands, outperformed their M1 baseline at most metrics, while the rest of the models did not. In fact, the accompaniment SIR of the M1 baseline was found to be statistically significantly better versus the M3 model. No conclusive results could be yielded from the comparison between models S1 and S2.

Discussion: The quantitiative results indicate that the Conv-TasNet architecture can benefit from the proposed multi-band technique. We assume that the separators, constrained in a smaller sub-space, learn to process the available information more efficiently, improving the separation performance. This can be backed by the fact that incorporating the features of an additional separator that processes the whole latent space (M3) worsens performance, leading to results close to the baseline. However, in terms of scalability of the technique to a larger number of bands, the performance of model M4 in comparison to M2 implies that using more separators, each being assigned a "narrower" latent space has diminishing returns.

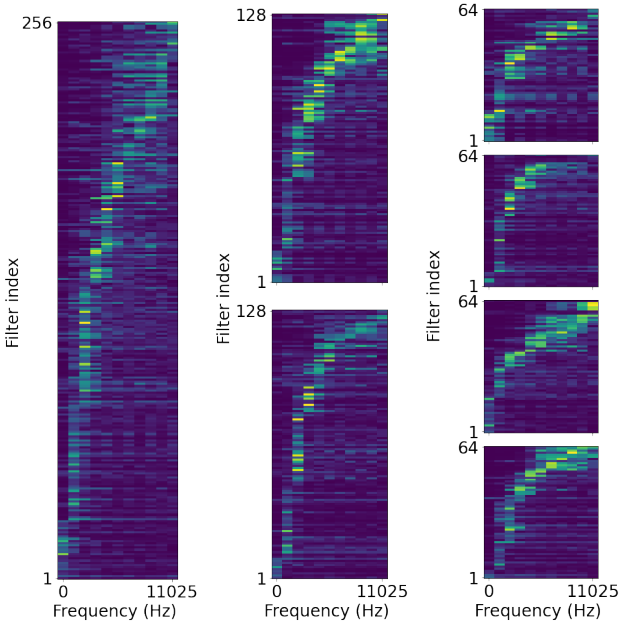

Fig. 3. Frequency domain representation of the encoder filters for models M1 (left), M2 (center), M4 (right). The sub-figures display the sub-space that is processed by the respective separator, in the case of multi-band models. The filters of each band are sorted in ascending order of base frequency.

Fig. 3 displays the frequency domain representation of the learned encoder filterbank for models M1, M2 and M4, For each model, the sub-figures match the sub-space of the latent space learned by the encoder that is processed by the respective separator. The filter responses are sorted by the filter's central frequencies in the frequency domain, in ascending frequency order. We can see that the single band/separator of the baseline model M1 forces the encoder to learn a latent space that covers all the frequencies, in a non-linear manner. For the M2 model, the two encoders also cover the same frequency range. Despite the appearance of small sub-bands with common characteristics in each encoder, such as the narrow low-frequency filters in the second band (bottom), the distribution of filters between the two encoders is mostly similar. However, for the model M4, the differences between the spanned sub-spaces of each encoder become more visible; for instance, the second band has converged to a sub-space containing more filters with lower central frequency and smaller bandwidth compared to the rest, while the bottom two bands contain visibly more high-frequency and high-bandwidth filters. These deviations are larger than the internal ones of the models M1 and M2, implying that the individual separators specialize in different regions of the spectrum. However, this behavior appears to be detrimental to the separation performance, indicating that

a larger latent space is required for each separator.

Regarding the M5 model, it is remarkable that it performed as well as M2, even though it had to adapt to a fixed latent space, optimized for the task of singing voice separation but trained independently. This might be an indication that the multi-band technique can work for models following the encoder-separator-decoder architecture with an arbitrary waveform-based front-end. However, the performance of the M6 model was slightly worse than M5, which implies that manually crafting bands in terms of spectral content from learnable filterbanks does not yield the performance gains it does in the case of STFT [8].

Finally, the multi-band technique didn't provide any boost to the model incorporating the stronger encoder (S2). Our explanation for this is that the dual nature of the extracted features, since they originate from both waveform and T-F domains, as well as the multi-layered structure of the encoder lead to a latent space structurally different than that of the Conv-TasNet, which is computed via a linear filterbank.

VI. CONCLUSION AND FUTURE WORK

In this work, we proposed a multi-band, multi-separator extension for models that follow the encoder-separator-decoder architecture, and investigated its application on the Conv-TasNet architecture. Our results in the task of Singing Voice Separation indicate that this approach can lead to improved performance compared to the baseline Conv-TasNet architecture, and is compatible with learned, waveform-based front-ends. As future work, we would like to validate the above claim by applying this technique in other waveform-based architectures [16] in the more general task of music source separation. Additionally, with recent research indicating the robustness of Digital Signal Processing - inspired filterbanks [23], we are interested in examining whether the latent spaces they generate can be used to perform a more sophisticated assignment of channels to bands and separators.

ACKNOWLEDGMENT

The authors wish to thank A. Zlatintsi (Nat. Tech. Univ. Athens) for the discussions regarding the content of the paper.

REFERENCES

[1] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18- A Corpus for Music Separation," https://doi.org/10.5281/zenodo.1117372, 2017.

[2] Z. Rafii, A. Liutkus, F. Stöter, S. Mimilakis, D. Fitzgerald, and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[3] F. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Proc. LVA/ICA 2018*, Guildford, UK, 2018.

[4] F. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A Reference Implementation for Music Source Separation," *Journal of Open Source Software*, vol. 4, p. 1667, 09 2019.

[5] S. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A Recurrent Encoder-Decoder Approach with Skip-filtering Connections for Monaural Singing Voice Separation," in *Proc. ICASSP 2018*, Calgary, Canada, 2018.

[6] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural Audio Source Separation Using Deep Convolutional Neural Networks," *Lecture Notes in Computer Science*, vol. 10169, pp. 258–266, 02 2017.

[7] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing Voice Separation with Deep U-Net Convolutional Networks," in *Proc. ISMIR 2017*, Suzhou, China, 2017.

[8] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation," in *Proc. IWAENC 2018*, Tokyo, Japan, 2018.

[9] N. Takahashi and Y. Mitsufuji, "Multi-scale Multi-band Densenets for Audio Source Separation," in *Proc. WASPAA 2017*, New Waltz, NY, USA, 2017.

[10] D. Griffin and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[11] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proc. ISMIR 2018*, Paris, France, 2018.

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[13] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," *arXiv preprint arXiv:1911.13254*, 2019.

[14] Z. Shi et al., "End-to-End Monaural Speech Separation with Multi-Scale Dynamic Weighted Gated Dilated Convolutional Pyramid Network," in *Proc. Interspeech 2019*, Graz, Austria, 2019.

[15] T. Nakamura and H. Saruwatari, "Time-Domain Audio Source Separation Based on Wave-U-Net Combined with Discrete Wavelet Transform," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[16] E. Nachmani, Y. Adi, and L. Wolf, "Voice Separation with an Unknown Number of Multiple Speakers," in *Proc. ICML 2020*, Vienna, Austria, 2020.

[17] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *Proc. ICASSP 2018*, Calgary, Canada, 2018.

[18] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient Long Sequence Modeling for Time-domain Single-channel Speech Separation," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[19] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning Extractors for Music Source Separation," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[20] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step Sound Source Separation: Training on Learned Latent Targets," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[21] A. Caillon and P. Esling, "RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

[22] Yu, Chengzhu et al., "DurIAN: Duration Informed Attention Network for Speech Synthesis," in *Proc. Interspeech 2020*, 2020.

[23] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[24] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An Empirical Study of Conv-TasNet," in *Proc. ICASSP 2020*, Barcelona, Spain, 2020.

[25] Z. Shi et al., "FurcaNet: An End-to-end Deep Gated Convolutional, Long Short-Term Memory, Deep Neural Networks for Single Channel Speech Separation," 2019.

[26] C. Lea, R. Vidal, A. Reiter, and G. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," in *Proc. ECCV 2016*, Amsterdam, the Netherlands, 2016.

[27] S . Uhlich et al., "Improving Music Source Separation Based on Deep Neural Networks through Data Augmentation and Network Blending," in *Proc. ICASSP 2017*, New Orleans, LA, USA, 2017.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–Half-Baked or Well Done?" in *Proc. ICASSP 2019*, Brighton, UK, 2019.