A MULTI-PHASE GAMMATONE FILTERBANK FOR SPEECH SEPARATION VIA TASNET

David Ditter and Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
david.ditter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

In this work, we investigate if the learned encoder of the end-to-end convolutional time domain audio separation network (Conv-TasNet) is the key to its recent success, or if the encoder can just as well be replaced by a deterministic hand-crafted filterbank. Motivated by the resemblance of the trained encoder of Conv-TasNet to auditory filterbanks, we propose to employ a deterministic gammatone filterbank. In contrast to a common gammatone filterbank, our filters are restricted to 2 ms length to allow for low-latency processing. Inspired by the encoder learned by Conv-TasNet, in addition to the logarithmically spaced filters, the proposed filterbank holds multiple gammatone filters at the same center frequency with varying phase shifts. We show that replacing the learned encoder with our proposed multi-phase gammatone filterbank (MP-GTF) even leads to a scale-invariant source-to-noise ratio (SI-SNR) improvement of 0.7 dB. Furthermore, in contrast to using the learned encoder we show that the number of filters can be reduced from 512 to 128 without loss of performance.

Index Terms— Speech Separation, Auditory Filterbank, End-To-End Learning, TasNet

1. INTRODUCTION

The introduction of end-to-end systems to the problem of monaural speech separation has led to significant performance gains in recent years. In contrast, deep learning approaches such as Deep Clustering [1], Permutation Invariant Training (PIT) [2, 3], Deep Attractor Networks [4] and Chimera++ [5] tackle the separation problem by transforming the input mixed speech into its Short-Time Fourier Transform (STFT) representation and then training a network to find an optimal multiplicative mask for each speaker in this domain. These STFT magnitude based approaches show a reasonable performance for the separation task but their performance is limited by the ideal mask calculation which typically does not include phase information.

These structural limitations are circumvented by end-to-end speech separation systems such as TasNet [6], Conv-TasNet [7] or FurcaNext [8, 9]. These systems introduce several changes to the STFT magnitude based approaches: First, the training loss is defined in the time domain instead of the STFT domain. Secondly, this time domain training loss allows for replacing the STFT analysis and synthesis stages by learned convolutional analysis and synthesis filterbanks. As in [6], we refer to the convolutional analysis filterbank as the *encoder* and the convolutional synthesis filterbank as the *decoder*. Thirdly, the encoder and decoder exhibit a significantly lower frame size than the STFT analysis-synthesis windows of state-of-the-art STFT-based approaches [1, 2, 3, 4, 5]. These reduced frame sizes allow for a strongly reduced algorithmic latency if the separation network does not employ look-ahead. And

finally, at least for the convolutional time domain audio separation network (Conv-TasNet) and FurcaNext, the separation section of the network is implemented as a temporal convolutional network with a bottleneck structure [10] instead of an architecture using Long Short-Term Memory (LSTM) layers. In [7], this modification has empirically shown to give better average results and to improve the robustness of these systems against time shifts of the input signal.

When replacing the deterministic STFT analysis-synthesis structure by a learned encoder-decoder structure, the following general question arises: Should we use a well-understood, deterministic encoder (analysis filterbank) which is based on signal processing principles and possibly motivated by perceptual features? Or should we let the network run free and find a data-driven signal encoding for the given problem all by itself? This question has very recently gained attention and was investigated in several research papers such as [11], [12] and [13].

On a theoretical level, there are good arguments for both choices. Advocating for a learned encoder, we can argue that we might obtain a better network after training if the network has a high degree of freedom and given that we have diverse and comprehensive training data. Advocating for a deterministic, hand-crafted encoder such as an STFT analysis or a gammatone filterbank, we can argue that a smaller number of trainable parameters limits the variance of the model and its ability to overfit. Hand-crafting the feature extraction allows for a regularization of the optimization problem. While this may potentially lead to a worse training optimum, the network may potentially generalize better, in particular if the training data is limited.

In this work, we propose to replace the learned encoder of the network by a deterministic gammatone filterbank motivated by auditory features. By this, we want to enlighten the discussion on learned versus deterministic signal encoding for the speech separation problem. We will furthermore show that replacing the learned encoder with our proposed multi-phase gammatone filterbank (MP-GTF) leads to an overall improvement from 15.4 dB to 16.1 dB measured in average scale-invariant source-to-noise ratio (SI-SNR) and allows to reduce the number of filters from 512 to 128 without performance loss.

We will first give an overview over our experimental framework in Section 2 and then go into details about our proposed MP-GTF for the encoder in Section 3. In Section 4 we present our results to then come to our conclusions in Section 5.

2. EXPERIMENTAL FRAMEWORK

2.1. Conv-TasNet

The architecture of the utilized neural network for speech separation is the Conv-TasNet architecture as presented in [7]. In Fig. 1 we show the overall structure of Conv-TasNet which consists of three
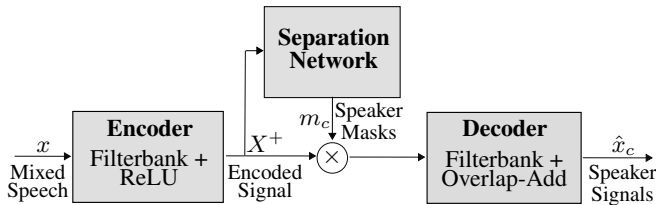
Fig. 1. Conv-TasNet [7] architecture. In this work we experiment with the encoder and decoder stage while the separation network parameters remain untouched.

Symbol	Description	Value
N	Number of filters in encoder / decoder	*varying*
L	Length of filters in samples	16
D	Frame shift in samples	8
B	Number of channels in bottleneck and residual paths' 1×1-conv blocks	256
H	Number of channels in convolutional blocks	512
P	Kernel size in convolutional blocks	3
X	Number of convolutional blocks in each repeat	8
R	Number of repeats	4

Table 1. Conv-TasNet hyper-parameter settings. The parameter N is varied in our experiments and does only affect the encoder and the decoder of the network. All other parameters remain fixed for all carried out experiments.

main structural elements, namely the encoder, the separation network and the decoder.

The *encoder* can be characterized as a filterbank with N filters of length L. The output signal $X(n,i) \in \mathbb{R}^{N \times \lfloor T/D \rfloor}$ of the n-th filter h_n^{Enc} for the mixed speech input signal $x \in \mathbb{R}^T$ is calculated as

$$X(n,i) = \sum_{l=0}^{L-1} x(iD+l)h_n^{\mathrm{Enc}}(L-l) \tag{1}$$

where i is the frame index, D is the frame shift and T is the length of the input signal. As described in [6, 7], we provide the separation network a non-negative signal. Thus, we eliminate the negative values using

$$X^+(n,i) = \max(0, X(n,i)). \tag{2}$$

which can be implemented by a Rectified Linear Unit (ReLU) layer. To account for this potential information loss, we propose to include phase inverted filters in the deterministic filterbank in Section 3.2.

The *separation network* which is explained in detail in [7] remains untouched for all our experiments with the hyper-parameters set fixed as shown in Table 1. We run the network in its non-causal configuration and use the global layer normalization method as detailed in [7]. At its output, the separation network provides a non-negative mask value $m_c(n,i) \in \mathbb{R}^+$ for the c-th speaker, all filters $n \in \{1, .., N\}$ and all time frames i. We obtain an estimate of each speaker in the encoded domain by the multiplication

$$X_c^+(n,i) = X^+(n,i)m_c(n,i). \tag{3}$$

In the *decoder* we reconstruct the estimated time domain signal of the c-th speaker $\hat{x}_c \in \mathbb{R}^T$ by first calculating intermediary sums for each filter weight index l and all time frames i as

$$x_c(l,i) = \sum_{n=0}^{N-1} X_c^+(n,i)h_{N-n}^{\mathrm{Dec}}(l) \tag{4}$$

and then executing the overlap-add operation as

$$\hat{x}_c(k) = \sum_{i=-\infty}^{\infty} x_c(k-iD,i) \tag{5}$$

where h_n^{Dec} holds the coefficients of the n-th filter of the decoder filterbank. In [7] the coefficients of h_n^{Enc} and h_n^{Dec} are initiliazed randomly and learned at training time. In this work we propose to replace the learned coefficients with the values of a deterministic filterbank. Also note that in [7] a matrix notation is used for description of the main structural elements of Conv-TasNet which simplifies the implementation of Conv-TasNet as a neural network. The notation as presented in this work is used to highlight signal processing properties of the encoder and decoder of Conv-TasNet.

2.2. Network configurations

For training and evaluating the network we use the WSJ0-MIX2 data set [1] which is commonly used for the speaker-independent monaural speech separation task. It consists of mixtures of two speakers mixed at power ratios in between 0 and 10 dB. The total amount of audio is roughly 20 hours in the training set, 5 hours in the validation set and 3 hours in the test set. For measuring the test set performance for a certain network configuration, we consider the training epoch where the validation set loss is minimal. The network training procedure is stopped when the validation set error has not improved within the last 10 epochs. All our experiments and evaluations are carried out at 8 kHz sampling rate.

2.3. Evaluation metrics

For performance comparison of all tested configurations and for use as a training objective we use the SI-SNR [6] of a speech signal calculated as

$$s_{\mathrm{target}} := \frac{\langle \hat{s}, s \rangle s}{||s||^2}, \tag{6}$$

$$e_{\mathrm{noise}} := \hat{s} - s_{\mathrm{target}}, \tag{7}$$

$$\mathrm{SI\text{-}SNR} := 10 \log_{10} \frac{||s_{\mathrm{target}}||^2}{||e_{\mathrm{noise}}||^2} \tag{8}$$

where $s, \hat{s} \in \mathbb{R}^T$ denote the clean and the estimated speech signals, $\langle \cdot, \cdot \rangle$ denotes the scalar product and $|| \cdot ||^2$ denotes the signal power. For the reported test set results we average the SI-SNR improvements over all 3000 mixtures.

3. FILTERBANK CONFIGURATIONS

The goal of this work is to investigate if the learned filterbanks of the end-to-end system Conv-TasNet is the key to its success, or if using a well defined deterministic filterbank performs equally well or even better. We have decided to choose modified auditory gammatone filterbanks for the deterministic filterbank for multiple reasons. Most importantly, an auditory gammatone filterbank (A-GTF) resembles the signal encoding in human auditory perception. As speech is a central sound source for humans, we assume that the human auditory system provides a signal representation which facilitates separation of different speakers. A second reason for using auditory gammatone
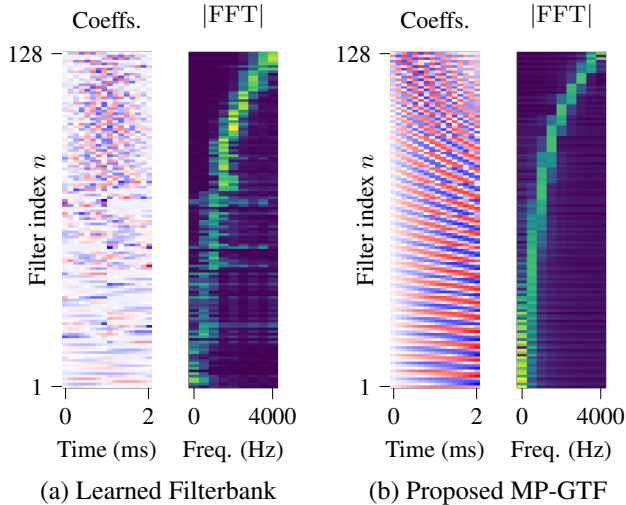
37

| Coeffs. | \|FFT\| | Coeffs. | \|FFT\| |

(a) Learned Filterbank   (b) Proposed MP-GTF

Fig. 2. (a) Time and frequency domain representation of all filters h_n^{Enc} for the encoder filterbank learned by Conv-TasNet with random initialization. (b) Time and frequency domain representation for all filters h_n^{Enc} of the proposed MP-GTF. The learned filterbank was sorted by the filter's peak values in the frequency domain and the number of filters is $N = 128$.

filterbanks is their non-linear spacing of center-frequencies, a structure that we also observe in the filterbanks learned by Conv-TasNet. In Figure 2 we plot the time domain and frequency domain representations of both the learned encoder filterbank and as well as for our proposed MP-GTF presented in Section 3.2. In the frequency representation of the learned filterbank, we can observe more filters that focus on the lower frequency regions than filters focusing on higher regions. We now first present the construction of a common A-GTF as presented in [14, 15] and then describe our modifications for usage in Conv-TasNet.

3.1. Auditory gammatone filterbank (A-GTF)

An A-GTF resembles the patterns of basilar membrane motion in the human auditory system. The filterbank consists of non-linearly spaced narrow-band filters γ_{f_c} with an increasing bandwidth over the filter's center frequency f_c. The impulse response $\gamma_{f_c}(t) \in \mathbb{R}$ of a real-valued gammatone filter is given by a gamma probability distribution function multiplied by a sinusoidal tone according to [14] as

$$\gamma(t) = a\, t^{(p-1)} e^{-2\pi b t} \cos\left(2\pi f_c t + \phi\right) \tag{9}$$

where f_c denotes the center frequency, ϕ the phase shift, a the amplitude, $t > 0$ the time in seconds, p the filter order and b the filter bandwidth parameter. [16] states that the filter order p measured for the human auditory system is in between 3 and 5.

The filter bandwidth parameter b and the center frequencies f_c are commonly determined using the concept of the equivalent rectangular bandwidth (ERB). The ERB estimates the bandwidth of the filters in the human auditory system using the simplification that each filter is a rectangular band-pass filter. The ERB has been found empirically as the following function over the center frequency f_c [17]:

$$\text{ERB}(f_c) = 24.7 + \frac{f_c}{9.265}. \tag{10}$$

The center frequencies of the human auditory filterbank are placed equally distant on the so called ERB scale which is derived by integrating $1/\text{ERB}(f_c)$ across frequency [15] resulting in

$$\text{ERB}_{\text{scale}}(f_{\text{Hz}}) = 9.265 \log(1 + \frac{f_{\text{Hz}}}{24.7 \times 9.265}). \tag{11}$$

where f_{Hz} denotes the frequency in Hertz. Thus, the relation between the acoustic frequency f_{Hz} and the ERB scale is given by a nonlinear logarithmic equation. A neighboring frequency with a distance of 1 on the ERB scale is then calculated as

$$f_{\text{next}} = \text{ERB}_{\text{scale}}^{-1}(\text{ERB}_{\text{scale}}(f_0) + 1) \tag{12}$$

where $\text{ERB}_{\text{scale}}^{-1}(f_{\text{ERB}})$ is the inverse of $\text{ERB}_{\text{scale}}(f_{\text{Hz}})$ with f_{ERB} denoting a frequency represented on the ERB scale.

To construct a complete A-GTF we typically define a lowest center frequency at around 50 Hz and define an upper limit for the filter center frequencies at around 8000 Hz. By using (12), we obtain all center frequencies of the filterbank by starting with the lowest center frequency and then iteratively going to higher frequencies until we reach the specified upper limit. For each center frequency we generate the time domain filters for the A-GTF via the impulse response function given in (9). As detailed in [17] the filters are normalized by their peak value in the freqency domain by varying the amplitude parameter a in (9) and the phase is set as $\phi = 0$.

3.2. Multi-phase gammatone filterbank (MP-GTF) [1]

To utilize a gammatone filterbank as a deterministic encoder of Conv-TasNet as described in Section 2, we propose several adaptations to the A-GTF. We name our proposed analysis filterbank for Conv-TasNet the *multi-phase gammatone filterbank (MP-GTF)*.

First, to keep the implementation close to original Conv-TasNet, we truncate the infinite impulse response (9) to a short length of 2 ms which corresponds to a length of 16 samples at 8 kHz sampling rate. This choice of a short filter length allows to keep the system latency low when using a causal separation network. Setting the filter order in between 3 and 5 as in the A-GTF leads to filters at low center frequencies which do not contain the peak amplitude of the impulse response (9). For this reason, in our modified MP-GTF we use the filter order of $p = 2$, which shifts the amplitude peaks towards the time origin and inside the incorporated time span of the gammatone. For this filter order, the bandwidth parameter for the gammatone at the center frequency f_c equates to $b = ERB(f_c)/1.57$ according to [17, Eq. 14].

Our second modification of the A-GTF for use in Conv-TasNet is to reduce the frequency range of the filter bank. The lowest center frequency of our proposed filterbank is chosen as 100 Hz while the upper range limit is set to 4000 Hz which is the Nyquist frequency for the utilized sampling rate of 8000 Hz. With the same construction method as for the A-GTF and the same spacing of 1 on the ERB scale, 24 center frequencies can be fitted within this limited frequency range.

As the third adaption and to meet the non-negativity constraint for the input to the separation network of Conv-TasNet, for each filter h_n^{Enc} in the encoder, we include the negative version $-h_n^{\text{Enc}}$ in our filterbank. Thus, we ensure that for each center frequency and each frame there is at least one filter that contains energy, if the original signal holds energy at this particular frequency. For a gammatone

---

[1]We provide Python code for construction of our proposed MP-GTF under https://github.com/sp-uhh/mp-gtf

38

Encoder	Decoder	N	SI-SNRi (dB)	
			Train	**Test**
Learned	Learned	512	19.3	15.4
MP-GTF	Learned	512	19.0	**15.9**
MP-GTF	MP-GTF Pseudo Inv.	512	18.1	15.4

Table 2. SI-SNR improvements on WSJ0-MIX2 training and test set for different configurations of the encoder and decoder of Conv-TasNet for $N = 512$ filters. Higher is better.

Encoder	Decoder	N	SI-SNRi (dB)
Learned	Learned	**512**	**15.4**
Learned	Learned	128	15.2
MP-GTF	Learned	512	15.9
MP-GTF	Learned	**128**	**16.1**
MP-GTF	Learned	64	15.4
MP-GTF	Learned	48	14.4

Table 3. SI-SNR improvements on WSJ0-MIX2 test set for different values of the number of filters N with a learned decoder.

filter as in (9) we can obtain this negative filter with a phase shift of π as $\phi^{\text{inv}} = \phi + \pi$.

As the last adaption, we propose a method to control the number of filters N in the encoder filterbank by introducing multiple filters with different phase shifts for the same center frequency. The minimum number of N based on our proposed construction is 48 due to the 24 center frequencies and the need to include the phase inverted version of each filter. If $N > 48$, we introduce multiple filters at each center frequency by choosing multiple values of ϕ. In total, we can only choose $N/2$ filters freely, as the other half is determined by the need for the phase inverted filters. This also implies that, based on our construction, N must be even. For a fixed center frequency f_c we introduce a total number of $\lfloor N/2/24 \rfloor$ phases where 24 is the number of center frequencies. As the division might hold a remainder, we distribute the remaining number of phase shifts to the lowest center frequencies. Once the number of phase shifts for a specific frequency is determined, we place the phase shifts equidistantly on the interval $[0, \pi)$ with the first phase shift being 0. We construct the first half of the filters at this center frequency by inserting these different values of ϕ into the gammatone impulse response (9). The other half of filters for this center frequency is generated by inserting all inverted phase shifts in the interval $[\pi, 2\pi)$ into the gammatone impulse response (9). In Figure 2, we plot the time and frequency domain representation of our proposed MP-GTF with $N = 128$.

4. RESULTS

In Table 2, we present the results of our experiments for different configurations of the encoder and decoder of Conv-TasNet with the number of filters fixed at $N = 512$. For the original configuration as in [7] we report a test set performance of $15.4\,\text{dB}$ where [7] reports a comparable performance of $15.6\,\text{dB}$. When replacing the encoder with our proposed deterministic MP-GTF and the decoder with the pseudo-inverse [18] of the MP-GTF, we reach a comparable performance of $15.4\,\text{dB}$. Table 2 also shows that the gap between the training and the test set is reduced from $3.9\,\text{dB}$ to $2.7\,\text{dB}$ in this configuration which suggests that a deterministic encoder and decoder combination is less prone to over-fitting. The optimal choice in terms of overall performance is to use our proposed MP-GTF in the encoder and to use a learned decoder which results in a performance of $15.9\,\text{dB}$ for $N = 512$. The decoder weights are initialized with the pseudo-inverse of the MP-GTF in this setting. A small number of reruns of all configurations has shown little variance for the test set performances.

In Table 3, we present the average test set performance for a varying number of filters N. Similar to [7], we found that for the original Conv-TasNet with learned filterbanks, the optimal performance is reached for $N = 512$ while lowering the number of filters to $N = 128$ slightly decreases the performance by a value of $0.2\,\text{dB}$. In contrast, when replacing the learned encoder with our proposed deterministic MP-GTF we find that we can lower the number of fil-

ters to $N = 128$ without performance loss, even resulting in the overall best model with an average SI-SNR performance of $16.1\,\text{dB}$. When setting $N = 48$, i.e. without redundant phase information in the encoder, the performance drops to $14.4\,\text{dB}$. This supports the hypothesis that amending the filters of the encoder with phase shifted versions is important for the performance gain of Conv-TasNet.

In addition to the values presented in the tables, we carried out performance analyses based on the difference of the median fundamental frequencies of the speakers within a mixture. As we have shown in [19], the median fundamental frequency difference is an important influencing factor to the performance of monaural speech separation systems and it is of importance to improve the performance especially for mixtures of similar speakers where the fundamental frequency difference is below $50\,\text{Hz}$. For our best model we were able to improve the performance in this important region of fundamental frequency difference by $0.9\,\text{dB}$ in contrast to Conv-TasNet with a learned filterbank.

Our findings imply that for the training data and architecture used for Conv-TasNet, the learned encoder is not critical for the performance gain of Conv-TasNet with respect to competing approaches. In contrast, a hand-crafted deterministic encoder filterbank performs even slightly better. In particular, it may reduce the gap between training and testing performance and also gives better results for difficult speaker pairs with similar fundamental frequencies, both of which indicates an increased robustness. Also, compared to the learned encoder, the number of filters N can be reduced without sacrificing performance.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed if the learned encoder is critical for the recent success of Conv-TasNet for monaural speech separation. We have shown that the overall system performance does not decrease when the learned encoder is replaced by our proposed deterministic multi-phase gammatone filterbank (MP-GTF). Instead, we can even measure a slight performance increase of $0.7\,\text{dB}$ in average SI-SNR. Furthermore with MP-GTF as the encoder filterbank, the number of filters can be reduced from 512 to 128 filters without compromising the overall performance. With the encoder filterbank set to the proposed MP-GTF and the decoder set as its pseudo-inverse, we showed that the overall performance is similar to the fully end-to-end system while over-fitting on the training data is reduced. In future work, it should be investigated if our findings hold for a broader range of speech processing tasks that can be tackled by neural networks such as monaural speech enhancement. Furthermore, we suggest to investigate if our proposed encoder leads to a more robust system by testing it with a larger variety of test data than given by the WSJ0-MIX data set. As a deterministic approach seems less prone to over-fitting, the proposed MP-GTF may provide for a more robust system in real-world applications.

39

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative Embeddings for Segmentation and Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35.

[2] D. Yu, M. Kolbaek, Z. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 241–245.

[3] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep Attractor Network for Single-Microphone Speaker Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 246–250.

[5] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 686–690.

[6] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 696–700.

[7] 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, Aug. 2019.

[8] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, S. Harada, and J. Han, "End-to-End Monaural Speech Separation with Multi-Scale Dynamic Weighted Gated Dilated Convolutional Pyramid Network," in *ISCA Interspeech*, Graz, Austria, Sept. 2019, pp. 4614–4618.

[9] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep Attention Gated Dilated Temporal Convolutional Networks with Intra-Parallel Convolutional Modules for End-to-End Monaural Speech Separation," in *ISCA Interspeech*, Graz, Austria, Sept. 2019, pp. 3183–3187.

[10] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1003–1012.

[11] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A Dissecting Approach," *arXiv preprint arXiv:1911.08895 [cs.SD]*, Nov. 2019.

[12] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-To-End Training of Time Domain Audio Separation and Recognition," *arXiv preprint arXiv:1912.08462 [cs, eess]*, Dec. 2019.

[13] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank Design for End-To-End Speech Separation," *arXiv preprint arXiv:1910.10400 [cs, eess]*, Oct. 2019.

[14] R. Patterson, Ian Nimmo-Smith, J. Holdsworth, and P. Rice, "An Efficient Auditory Filterbank based on the Gammatone Function," *MRC Applied Psych. Unit*, 1988.

[15] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[16] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex Sounds and Auditory Images," in *Auditory Physiology and Perception*, pp. 429–446. Elsevier, 1992.

[17] V. Hohmann, "Frequency Analysis and Synthesis using a Gammatone Filterbank," *Acta Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[18] R. Penrose, "A Generalized Inverse for Matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, July 1955.

[19] D. Ditter and T. Gerkmann, "Influence of Speaker-Specific Parameters on Speech Separation Systems," in *ISCA Interspeech*, Graz, Austria, Sept. 2019, pp. 4584–4588.