Filterbank design for end-to-end speech separation

Manuel Pariente, Samuele Cornell, Antoine Deleforge, Emmanuel Vincent

FILTERBANK DESIGN FOR END-TO-END SPEECH SEPARATION

Manuel Pariente [1]*, Samuele Cornell* [2]*, Antoine Deleforge* [1]*, Emmanuel Vincent* [1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Department of Information Engineering, Università Politecnica delle Marche, Italy

ABSTRACT

Single-channel speech separation has recently made great progress thanks to learned filterbanks as used in ConvTasNet. In parallel, parameterized filterbanks have been proposed for speaker recognition where only center frequencies and bandwidths are learned. In this work, we extend real-valued learned and parameterized filterbanks into complex-valued analytic filterbanks and define a set of corresponding representations and masking strategies. We evaluate these filterbanks on a newly released noisy speech separation dataset (WHAM). The results show that the proposed analytic learned filterbank consistently outperforms the real-valued filterbank of ConvTasNet. Also, we validate the use of parameterized filterbanks and show that complex-valued representations and masks are beneficial in all conditions. Finally, we show that the STFT achieves its best performance for 2 ms windows.

Index Terms— Speech separation, filterbank design.

1. INTRODUCTION

Be it for speech intelligibility or automatic speech recognition, speech processing applications need effective speech separation in clean and noisy recording conditions. Single-channel speaker-independent speech separation has recently seen great progress in clean recording conditions. A wide variety of deep learning methods have been introduced [1–10] and compared on the wsj0-2mix benchmark introduced in [1]. All these methods rely on a neural network to estimate the time-frequency mask associated with each source.

Crucially, the time-frequency transform must allow both signal analysis and resynthesis. It can be either fixed, such as the short time Fourier transform (STFT) [1–6] and its inverse and the Mel [11] or gammatone [12] filterbanks, or learned

jointly with the masking network [7–10]. While learned representations have been shown to be undeniably superior to the STFT for speech separation in clean recording conditions [7–9], their impact in the presence of noise has been lesser studied. In fact, the authors of [13] introduce a noisy extension of wsj0-2mix, WHAM, on which initial results indicate that the advantage of learned representations reduces as noise is introduced, suggesting that learning from the raw waveform might be harder in noisy conditions. In parallel, parameterized kernel-based filterbanks have been introduced as a front-end for speech and speaker recognition [14, 15]. The underlying idea is to restrict the filters to a certain family of functions and jointly learn their parameters with the network. These filterbanks are meant for signal analysis only, though.

In this paper, we define suitable parameterized filters for analysis-synthesis. Compared with fixed STFT filters, the proposed filters offer more flexibility and diversity thanks to their adaptive center frequency and bandwidth. Conversely, their parameterized form offers fewer parameters to learn and better interpretability compared with their learned counterparts. To do so, we extend the parameterized filters introduced in [14] to complex-valued analytic filters, thus enabling perfect synthesis via overlap-add and building shift invariance, a desirable property for time-frequency representations. We then propose a similar analytic extension for learned filters. Finally, we evaluate the performance of these analysis-synthesis filterbanks in a unified framework, as a function of window size in both clean and noisy scenarios.

We present the general framework for speech separation and the proposed filterbanks in Section 2. We describe the experimental setup in Section 3 and the results in Section 4. We conclude in Section 5.

2. MODEL

Single-channel speech separation is the task of retrieving individual speech sources from a mixture, optionally in the presence of noise. The observed signal $x(t)$ is described as

$$x(t) = \sum_{i=1}^{C} s_i(t) + n(t), \tag{1}$$

where C is the number of sources, $\{s_i(t)\}_{i=1..C}$ are the individual source signals and $n(t)$ is additive noise. The task is then to produce accurate estimates $\hat{s}_i(t)$ of each $s_i(t)$.

2.1. General framework

Most state-of-the-art speech separation methods can be described using an encoding-masking-decoding framework. An encoder transforms the time-domain signal by convolving every signal frame indexed by $k \in \{0, ..., K-1\}$ with a bank of N analysis filters $\{u_n(t)\}_{n=0..N-1}$ of length L:[1]

$$\mathbf{X}(k,n) = \sum_{t=0}^{L-1} x(t+kH)u_n(t), \quad n \in \{0, ..., N-1\}, \quad (2)$$

where H is the *hop size*. After an optional non-linearity $\mathcal{G}$, $\mathbf{X}$ is then fed to the masking network $\mathcal{MN}$:

$$\mathcal{MN}(\mathcal{G}(\mathbf{X})) = [\mathbf{M}_1, ..., \mathbf{M}_C]. \quad (3)$$

Each estimated mask $\mathbf{M}_i$ is multiplied with the input to obtain the estimated representation of source i:

$$\mathbf{Y}_i = \mathcal{G}(\mathbf{X}) \odot \mathbf{M}_i, \quad i \in \{1, ..., C\}, \quad (4)$$

with $\odot$ denoting point-wise multiplication. The decoder maps each $\mathbf{Y}_i$ to the time domain by transposed convolution with a bank of N *synthesis* filters $\{v_n(t)\}_{n=0..N-1}$ of length L:

$$\hat{s}_i(t) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \mathbf{Y}_i(k,n)v_n(t-kH). \quad (5)$$

The analysis and synthesis filters fall into three categories: *free*, *parameterized*, or *fixed*. In [7–9], the filters are free: all weights $\{u_n(t)\}$ and $\{v_n(t)\}$ are jointly learned with the masking network. Parameterized filters belong to a family of filters, whose parameters θ are jointly learned with the network instead [14, 15]. For instance, the filters in [14] are defined as the difference between two low-pass filters with cutoff frequencies f_1 and f_2:

$$u_n(t;\theta) = 2f_2 \operatorname{sinc}(2\pi f_2 n) - 2f_1 \operatorname{sinc}(2\pi f_1 n)$$
$$= 2f_w \operatorname{sinc}(2\pi f_w n) \cos(2\pi f_c n), \quad (6)$$

where $\theta = \{f_1, f_2\}$, $f_w = f_2 - f_1$, and $f_c = (f_1 + f_2)/2$. All filters drawn from this family are even funtions, thus making it unsuitable for resynthesis. Finally, fixed filters represent handcrafted transforms such as the STFT [1], gammatone [12] or Mel [11] filters. In the case of the STFT:

$$u_n(t) = h_a(t)e^{-2j\pi n/N} \text{ and } v_n(t) = h_s(t)e^{2j\pi n/N}, \quad (7)$$

with h_a and h_s the analysis and synthesis windows.

---

[1]Mathematically, this is a correlation rather than a convolution.

A desirable property of time-frequency representations is *shift invariance*, i.e., invariance to small delays in the time domain. *Analytic filters* [16] have this property. Namely, the modulus of the convolution between a real-valued signal and an analytic filter is the envelope of that signal in the frequency band defined by the filter. The STFT filters (7) are examples of such analytic filters, and the magnitude of the STFT is the corresponding shift-invariant representation. Given any real-valued filter $u(t) \in \mathbb{R}^{1 \times L}$, a corresponding analytic filter $u_{\text{analytic}}(t)$ can be obtained as

$$u_{\text{analytic}}(t) = u(t) + j\mathcal{H}[u(t)] \quad (8)$$

where $\mathcal{H}$ denotes the Hilbert transform which imparts a $-\pi/2$ phase shift to each positive frequency component. In the following, we detail the proposed analytic expansion of both parameterized and free filters.

2.2. Proposed analytic filterbanks

We define parameterized analytic analysis filters u_n as

$$u_n(t;\theta) = 2f_w \operatorname{sinc}(2\pi f_w t)(\cos(2\pi f_c t) - j \sin(2\pi f_c t))$$
$$= 2f_w \operatorname{sinc}(2\pi f_w t)e^{-2j\pi f_c t}. \quad (9)$$

This complements the original family of even filters (6) with odd ones. The new family $\{u_n\}_{n=1..N}$ can form a complete basis of the signal space, and each filter is *analytic* so that

$$\Im(u_n(t;\theta)) = \mathcal{H}[\Re(u_n(t,\theta))]. \quad (10)$$

The corresponding family of synthesis filters is defined as:

$$v_n(t;\phi) = 2g_n f_w \operatorname{sinc}(2\pi f_w t)e^{2j\pi f_c t}, \quad (11)$$

where $\phi = \{f_1, f_2, g\}$, and g_n is a gain parameter learned to improve resynthesis. Finally, each filter is multiplied by a Hamming window of size L, as in [14, 15].

Similarly, in the case of free filters, we propose to ensure that the learned filters are analytic by parameterizing them by their real part and computing the corresponding analytic filter via (8) during the forward pass of the network. This is applied to both analysis and synthesis filters.

In the following, analytic parameterized and free filterbanks are respectively denoted as *free+$\mathcal{H}$* and *param+$\mathcal{H}$*.

2.3. Network inputs and output masks

Analytic filterbanks can be viewed either as a set of N complex filters or as $2N$ real filters. This opens different possibilities for the inputs and outputs of the masking network. We consider three possibilities for the input representation: the modulus of $\mathbf{X}$ (*Mag*), its real and imaginary parts (*Re+Im*) or a concatenation of both (*Mag+Re+Im*). The masks can be applied to the modulus of $\mathbf{X}$ (*Mag*) or to its real and imaginary parts using either a complex-valued product with $\mathbf{M}_k \in \mathbb{C}^N$ (*Compl*) or a real-valued product with $\mathbf{M}_k \in \mathbb{R}^{2N}$ (*Re+Im*).

2.4. Masking network

The masking network is chosen to be the time-domain Convolutional Network (TCN) in [9]. It comprises R convolutional blocks, each consisting of X 1-D dilated convolutional layers with exponentially increasing dilation factor. In [9], the best system used $R = 3$ and $X = 8$. For most of our experiments, we use a lighter network (*Light TCN*) with $R = 2$ and $X = 6$ to reduce training time. The systems achieving the best performances with the light TCN are then retrained using the larger network (*Full TCN*). The hop size is set to $H = L/2$. No non-linearity $\mathcal{G}$ is applied to the inputs and the masks are estimated using ReLU as the activation function[2].

3. EXPERIMENTAL PROCEDURE

3.1. Dataset

The systems are evaluated on clean (wsj0-2mix [1]) and noisy (WHAM [13]) two-speaker mixtures created with the scripts in [17]. In the clean condition, a 30 h training set and a 10 h validation set are generated by mixing randomly selected utterances from different speakers in the Wall Street Journal (WSJ) training set si_tr_s at random signal-to-noise (SNR) ratios between 0 and 5 dB. Noisy datasets are then created by mixing noiseless mixtures with noise samples at SNRs between -3 and 6 dB with respect to the loudest speaker. For both conditions, a 5 h evaluation set is designed similarly with different speakers and noise samples. All Light TCN experiments are conducted with a sampling rate of 8 kHz.

3.2. Training and evaluation setup

Training is performed on 4 s segments using the permutation-invariant [2,4] scale-invariant source-to-distorsion ratio (SI-SDR) [9,18] as the training objective. For Light TCN, Adam [19] with an initial learning rate of 1.10^{-3} is used as the optimizer. Learning rate halving and early stopping are applied based on validation performance. The best models are retrained with the Full TCN architecture using rectified Adam [20] with look ahead [21]. Mean SI-SDR improvement (SI-SDR$_i$) is reported for all models on their respective test sets.

4. RESULTS

4.1. Light TCN experiments

Analycity of parameterized and free filterbanks: We first evaluate the role of analycity in our parameterized filterbank

---

[2]Our implementation can be found at github.com/mpariente/AsSteroid.

---

(9). We consider $N = 512$ filters, with *Mag+Re+Im* input and *Mag* mask. To compensate for the greater number of filters, we set $N = 1536$ for the non-analytic filterbank (6). Table 1 shows that the original filterbank is unsuitable for analysis-synthesis for any window size and that the proposed analytic extension overcomes this issue.

Window size (ms)	2	5	10	25	50
Param.	2.3	1.0	0.6	-0.8	-2.7
Param.(3x filters)	2.3	1.2	0.7	-0.7	-2.7
Param.$+\mathcal{H}$	**11.8**	**11.6**	**9.1**	**7.3**	**4.0**

Table 1. SI-SDR$_i$ (dB) as a function of window size for parametric filterbanks in clean conditions.

The results of a similar experiment for the free filterbanks in both clean and noisy conditions are shown in Fig. 1. While analytic extension of free filters doesn't hurt performance for short windows, we can gain up to 2 dB for larger windows. Also note that tripling the number of free filters doesn't match the gain brought by analycity.
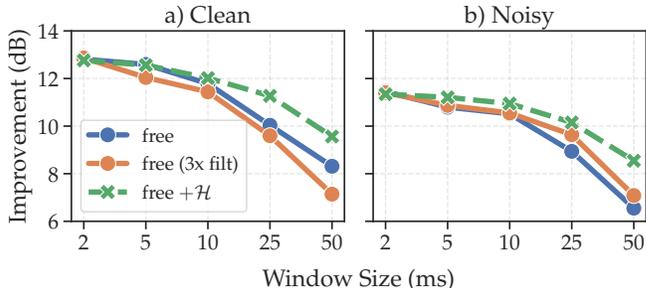

Fig. 1. SI-SDR$_i$ as a function of window size for free filterbanks in clean and noisy conditions.

Masking strategies: Next, we evaluate the choice of the masking strategy for *Mag+Re+Im* input. Table 2 shows that this has very little and moderate impact on the free$+\mathcal{H}$ and param.$+\mathcal{H}$ filterbanks, respectively. For the STFT, the gap between *Mag* and *Re+Im* masks suggest that phase modeling is indeed necessary for good separation with small windows.

Filterbank	Clean			Noisy		
	Mag	Compl	Re+Im	Mag	Compl	Re+Im
Free$+\mathcal{H}$	12.7	**12.8**	**12.8**	11.0	**11.3**	11.0
Param.$+\mathcal{H}$	11.8	12.2	**12.5**	10.5	**10.6**	10.1
STFT	9.8	10.5	**10.9**	9.4	9.4	**9.9**

Table 2. SI-SDR$_i$ (dB) as a function of the mask type for each analytic filterbank in clean conditions, with $L = 16$.

Input representations: Next, we evaluate the impact of the input representation as a function of window size. The results are plotted in Fig. 2 for the *Re+Im* mask. In the case of
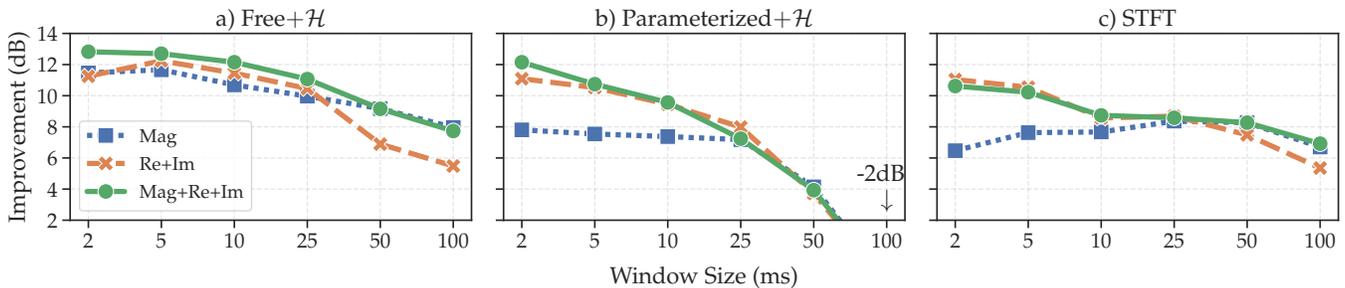
Fig. 2. SI-SDR$_i$ for different inputs to the network as a function of window size for a) free+$\mathcal{H}$, b) param.+$\mathcal{H}$ and c) STFT.

free+$\mathcal{H}$, the shift-invariant representation *Mag* helps only for large windows. For param.+$\mathcal{H}$, the *Re+Im* representation is sufficient for all window sizes. Finally, for the STFT, the *Mag* and *Re+Im* inputs complement each other so that maximum performance is reached with *Re+Im* input for small windows, when phase modeling is necessary, and with *Mag* for larger windows when amplitude modeling is sufficient.
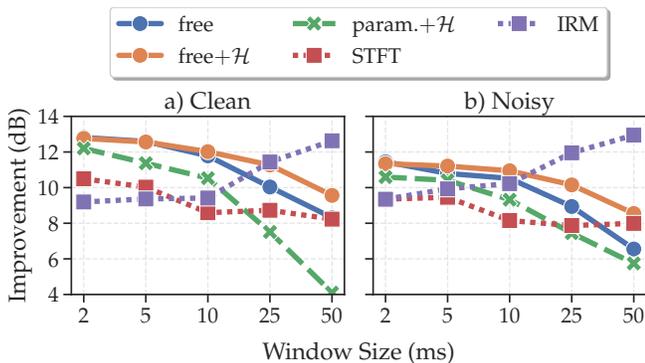

Fig. 3. SI-SDR$_i$ as a function of window size for all filterbanks and for the ideal ratio mask (IRM).

Filterbank choice: Finally, we compare the original free filterbank and all analytic filterbanks in Fig. 3, in both clean and noisy conditions. We use *Mag+Re+Im* input and *Re+Im* masking for all methods based on analytic filters.

The take-away messages of the above experiments are as follows. First, parameterized filters as in [14] are unsuitable for separation. The proposed analytic extension addresses this issue but performance decreases as the window size increases. Second, analytic extension of learned filters stabilizes performance for large windows. Third, combining complex inputs and masks for small windows brings the best results for all analytic filterbanks. Interestingly, this also holds for the STFT.

4.2. Full TCN experiment

We retrained the two best models, i.e., free with $L = 16$ (a.k.a. Conv-TasNet) and free+$\mathcal{H}$ with $L = 16$, with the full TCN in clean and noisy conditions and for both the 8 kHz *min*

and 16 kHz *max* versions of the dataset. The results are reported in Table 3 along with the chimera++ [6] results in [13]. Compared to Conv-TasNet, the proposed analytic extension improves the results in all tested conditions by up to 0.7 dB, showing that shift-invariant representations can benefit Conv-TasNet's TCN even for small windows.

Model	Dataset	separate-clean	separate-noisy
chimera++ [6]	8kHz min	11.0	9.9
Conv-TasNet [9][3]	8kHz min	15.1	12.7
Free+$\mathcal{H}$	8kHz min	**15.8**	**12.9**
chimera++ [6]	16kHz max	9.6	10.2
Conv-TasNet [9]	16kHz max	13.6	13.3
Free+$\mathcal{H}$	16kHz max	**14.0**	**14.0**

Table 3. SI-SDR$_i$ (dB) comparison between the proposed analytic free filterbank and previously proposed models.

5. CONCLUSION

In this paper, we defined analytic extensions of both parameterized and free filterbanks. The resulting filterbanks are more interpretable, and perform equally well or better than their real-valued counterparts for speech separation in clean or noisy conditions. Final evaluation with the most expressive TCN from [9] showed that the proposed analytic free filterbank consistently outperforms the non-analytic free filterbank employed in ConvTasNet and provides the best performance overall. Also, to the best of our knowledge, this is the first time parameterized filterbanks are used in an end-to-end speech separation framework. Although they don't perform as well as free filterbanks, we argue that a better design could bridge this gap. Finally, we showed that complex-valued inputs and masks can be beneficial for separation with short windows for all filterbanks. In particular, the implicit phase modeling capability of ConvTasNet's TCN applies to the STFT, which achieves its best performance for 2 ms windows.

---

[3]Our implementation.

6. REFERENCES

[1] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2016.

[2] D. Yu, M. Kolbk, 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)*, March 2017, pp. 241–245.

[3] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[4] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[5] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017.

[6] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 686–690.

[7] Yi Luo and Nima Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018.

[8] Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu, Jiqing Han, and Anyan Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation," in *Proc. Interspeech*, 2019, pp. 3183–3187.

[9] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] Gene-Ping Yang, Chao-I Tuan, Hung-Yi Lee, and Lin shan Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," in *Proc. Interspeech*, 2019, pp. 1363–1367.

[11] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Aug. 2015.

[12] Thibaud Necciari, Nicki Holighaus, Peter Balazs, Zdeněk Průša, Piotr Majdak, and Olivier Derrien, "Audlet filter banks: a versatile analysis/synthesis framework using auditory frequency scales," *Applied Sciences*, vol. 8, no. 1, pp. 96, 2018.

[13] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "WHAM!: extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[14] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[15] Erfan Loweimi, Peter Bell, and Steve Renals, "On Learning Interpretable CNNs with Parametric Modulated Kernel-Based Filters," in *Proc. Interspeech*, 2019, pp. 3480–3484.

[16] J. L. Flanagan, "Parametric coding of speech spectra," *The Journal of the Acoustical Society of America*, vol. 68, no. 2, pp. 412–419, 1980.

[17] "Scripts to generate the wsj0 hipster ambient mixtures dataset," http://wham.whisper.ai/.

[18] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 626–630.

[19] Diederik P. Kingma and Jimmy Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," 2019.

[21] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba, "Lookahead optimizer: k steps forward, 1 step back," 2019.