Interspeech 2022
18-22 September 2022, Incheon, Korea

Learnable Sparse Filterbank for Speaker Verification

Junyi Peng[1], Rongzhi Gu[2], Ladislav Mošner[1], Oldřich Plchot[1], Lukáš Burget[1], Jan Černocký[1]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]Peking University, ECE, China
pengjy@fit.vut.cz

Abstract

Recently, feature extraction with learnable filters was extensively investigated with speaker verification systems, with filters learned both in time- and frequency-domains. Most of the learned schemes however end up with filters close to their initialization (e.g. Mel filterbank) or filters strongly limited by their constraints. In this paper, we propose a novel learnable sparse filterbank, named LearnSF, by exclusively optimizing the sparsity of the filterbank, that does not explicitly constrain the filters to follow pre-defined distribution. After standard pre-processing (STFT and square of the magnitude spectrum), the learnable sparse filterbank is employed, with its normalized outputs fed into a neural network predicting the speaker identity. We evaluated the performance of the proposed approach on both VoxCeleb and CNCeleb datasets. The experimental results demonstrate the effectiveness of the proposed LearnSF compared to both widely-used acoustic features and existing parameterized learnable front-ends.

Index Terms: learnable filter, sparse filtering, sparsity, speaker verification

1. Introduction

In recent years, since deep learning has shown its remarkable success in speech modeling, more researchers focus on building deep structures [1, 2] or investigating effective objective functions [3, 4, 5] to extract discriminant speaker representations. Most of these approaches take hand-crafted acoustic features as input (e.g. Gammatones, log Mel filterbank (Mel-FBank), Mel frequency cepstral coefficients (MFCC)), which are inspired by the human auditory perception mechanism. Even though such features provide good performance, nowadays's neural approaches are powerful enough to re-train also the front ends and there is no reason to believe that hand-crafted features with fixed extraction parameters are optimal for all speech-related tasks.

To challenge those hand-crafted acoustic features, there have been several attempts on designing a learnable front-end for speaker embedding extractor [6, 7, 8]. Two main approaches exist: In the signal-based approach, the first block of those models is expected to model the vocal tract-related characteristics directly from the raw waveform or spectral features, the essence of which is a set of learnable filters [6, 8] that are jointly optimized with the following speaker embedding extractor. In [9], a modified convolutional layer composed of parameterized band-pass filters, named SincNet, is integrated to speaker model to extract speaker embedding from the raw waveform. Compared to traditional CNN, SincNet takes the advantages of a parametric model: higher interpretability and fewer parameters [10]. Since SincNet only focuses on the magnitude domain while ignoring the importance of phase, in [11], a bank of complex-valued time-domain filters, which functions as a proxy for learnable Mel-frequency spectral coefficients, is investigated for end-to-end phone recognition. The experimental results on the TIMIT dataset show superior performance over widely-used acoustic features (e.g. Mel-FBanks, MFCCs). Sharing a similar idea, [12] explores an interpretable complex-valued exponential filter (IC filter) for time-domain speaker verification, which enforces each filter to follow trigonometric functions parameterized by its center frequency. More recently, in [13], a learnable front-end (LEAF) for audio classification has been proposed, which consists of a complex-valued Gabor filterbank [11], Gaussian low-pass filter, and smoothed per-channel energy nomination [14]. The system reaches state-of-the-art performance on the AudioSet benchmark with a slight advantage over Mel-FBank features.

Secondly, learnable filters have also been developed in the frequency domain: [15] implements an unconstrained learnable filterbank, which provides frequency clues for the neural network to reduce the training loss of a frequency prediction task. In [16], a non-negative restricted learnable filterbank is leveraged to detect speaker-related information from the magnitude spectrum. Through joint optimization with the following classifier, the learned representation is more robust than manually designed features in the field of anti-spoofing.

Generally, although features extracted by the aforementioned learnable filters have shown competitive performances compared to hand-crafted acoustic features, their advantages are very limited: with such strong prior signal-processing-related constraints on filters' distributions or shapes, each filter has only a few learnable parameters. Hence it results in a clear time-frequency structure and high interpretability of learned filters. However, when analyzing the frequency responses, most of them just re-learn the Mel-scale. This might be due to strong constraints on the filter shapes and Mel-scale initialization. This combination endows the front-end with a fair ability to extract discriminative features at the beginning of the training process, while updating only slowly in the following optimization.

In this paper, we address the strong constraints of learnable front-end for speaker verification, by a learnable sparse filterbank (LearnSF). Given speech, only a small fraction of time-frequency patterns are relevant to speaker identity. In sparse filtering [17], this could be achieved by activating a few frequency components and reducing others to zero while maintaining diversity between filters. We start the processing by windowing and short-term Fourier transform (STFT), convert-

The work was partly supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, and Czech Ministry of Education, Youth and Sports from project no. LTAIN19087 "Multi-linguality in speech technologies" and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "e-Infrastructure CZ – LM2018140".

10.21437/Interspeech.2022-11309

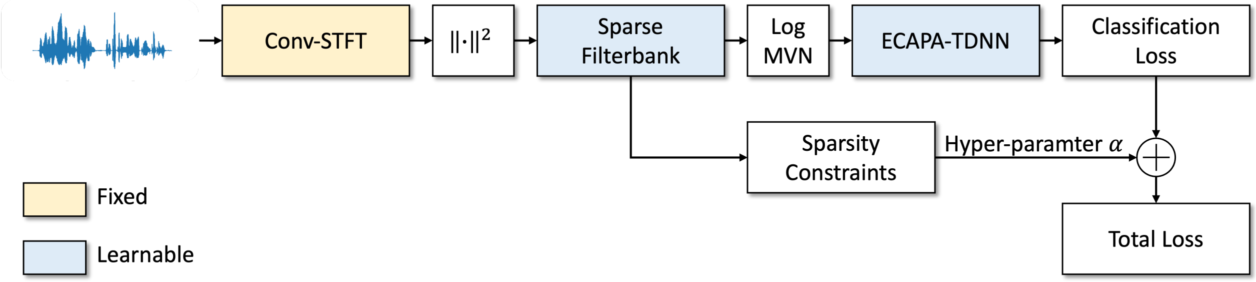


Figure 1: *Proposed LearnSF speaker embedding extractor. The yellow box denotes all coefficients are fixed, while the parameters in blue box can be updated by the training.*

ing speech into time-frequency representation. We convert the magnitude spectrum to power spectrum, which forms the learnable filter's input implemented as a standard matrix operation over the power spectrogram. The bank of filters is initialized by Mel-scale and later jointly optimized with the rest of the neural network – enforcing the sparsity constraints helps to define frequency regions that are relevant to the SV task while preventing degenrate solutions. Next, the filtered time-frequency features are passed through a logarithmic compression, and mean variance normalization. Furthermore these normalized features are fed into a neural network to predict the speaker identity. The experimental results on VoxCeleb and CNCeleb show that the proposed LearnSF outperforms the widely-used acoustic features such as Mel-FBanks, as well as different parameterized learnable filters based SV systems. Additionally, we show that our front-end generalizes better in cross-lingual scenario than hand-crafted acoustic features.

2. Learnable Sparse Filterbank

Figure 1 shows the overall scheme of LearnSF-based SV system. We start by standard windowing, STFT and power of two, to produce a power spectrogram of speech[1]. Then, the learnable filterbank follows; inspired by [17], an objective function is used to optimize the sparsity of the learned filterbanks. Different constraints, introduction of sparsity to filter learning, and different initialization methods are discussed in the following sub-sections. Finally, sequentially, the learned features are post-processed by logarithmic compression and mean variance normalization (MVN).

2.1. Formalism

The power spectrogram is defined as $\mathbf{S} \in \mathbb{R}^{N\times F}$, where N is the number of frames, and F is the number of STFT bins from 0 to half of the sampling frequency. The bank of K learnable filters applied in frequency domain is denoted as $\mathbf{V} \in \mathbb{R}^{F\times K}$, with individual filters $\mathbf{v}_{1\dots K}(\mathbf{v}_k \in \mathbb{R}^{F\times 1})$. The corresponding time-frequency output $\mathbf{O} \in \mathbb{R}^{N\times K}$ is given by:

$$\mathbf{O} = \mathbf{S}\mathbf{V}, \tag{1}$$

where F is also the length of learnable filters. In the following, we will denote the output of k-th filter at n-th frame as $O_{n,k}$.

2.1.1. Vanilla filters

In our early experiments, we employed a set of filters using random initialization without any constraints on parameters to model the power spectrum. The filters are expected to capture meaningful frequency bands (see Fig. 2 (b)), however, the training process of the whole SV system is unstable and the final performance (see section 3.4) is much worse than Mel-FBank based system. Also, it is not possible to ensure that the filterbank results will be non-negative. Therefore, compression by logarithm and mean- and variance-normalization (MVN) can not be applied.

2.1.2. Vanilla filters initialized by Mel-scale

Motivated by the success of learnable filters initialized by Mel-scale in automatic speech recognition (ASR) field [9, 19], we introduce Mel-scale to initialize the filters. As shown in Fig 2 (c), the Mel-scale initialization makes the learned filters have a clearer structure and a slight performance boost is achieved. However, the filters might still have negative gains, and, as in the previous experiments, log compression and MVN are not applicable.

2.1.3. Normalized filter

As [20, 21] point out that non-negative and band-limited constraints play an important role in designing filterbank for magnitude inputs. Therefore, we adopt l2-normalization to the parameters of each filter as:

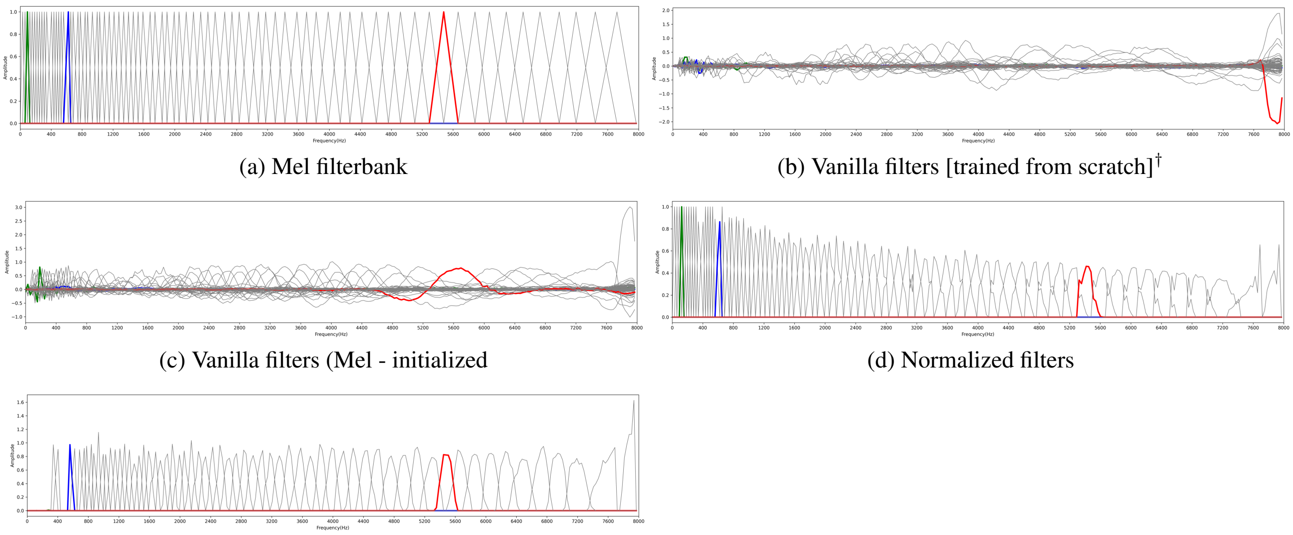
$$\hat{\mathbf{v}}_k = abs\left(\frac{\mathbf{v}_k}{\sum_{f=1}^{F}||\mathbf{v}_k[f]||^2}\right), \tag{2}$$

where $abs(\cdot)$ denotes taking the absolute value. This operation ensures positive-valued coefficients of the filterbank. Note that $\sum_{f=1}^{F}||\hat{\mathbf{v}}_k[f]||^2 = 1$ suggests that each filter has the same total gain, as also shown in Fig 2 (d). This modification leads to a significant improvement. However, the degree of freedom of filter coefficients is still high, and they might be prone to overfitting in the training process.

2.1.4. Sparse filter

To restrict the coefficients of learnable filters and boost performance, we introduce sparsity to the filtering, instead of forcing filters to follow some pre-defined function (e.g. Gaussian function [11]) or keep orthogonality (e.g. DFT coefficients). By sparsity constraints and joint training with neural network predicting the speaker identity, we expect to obtain filters capable of capturing speaker-related frequency patterns while staying

[1] In our implementation, to speed up the propagation and enable on-the-fly wave augmentation, as described in our previous work [12, 18], we define a 1D convolution layer, named Conv-STFT, as a combination of window function and DFT kernels, for production of the magnitude spectrogram



(a) Mel filterbank

(b) Vanilla filters [trained from scratch]†

(c) Vanilla filters (Mel - initialized

(d) Normalized filters

(e) Sparse filters

Figure 2: *Frequency responses of different filters at convergence. Except for Vanilla filters*†*, all other filters (vanilla filters, normalized filter, Gaussian filters and sparse filters) are initialized by a Mel scale. We highlight three among 80 filters: the 5th (green), the 20th (blue) and the 70th (red).*

insensitive to other variations. This can be achieved with learnable filters where only a few coefficients per-filter are active (i.e. non-zero) while all the others are close to 0. Intuitively, this kind of inner-filter sparsity can be directly measured by l_p-norm (i.e. $|\mathbf{v}_k|_p = \sqrt[p]{\sum_{f=1}^{F} \|v_k[f]\|^p}$) and constrained via an objective function computed as:

$$L_{direct} = \frac{1}{K}\sum_{k=1}^{K} |\mathbf{v}_k|_p, \tag{3}$$

where $\mathbf{v}_k$ is k-th filter. However, such a simple constraint could only ensure the sparsity of a single filter, while the diversity across filters is not guaranteed. Inspired by [17], we ensure diversity of the filter-bank by penalizing the outputs of the filters rather than their coefficients. This cross-filter constraint is expected to prevent degenerate scenarios in which the same patterns are modeled repeatedly. At first, we normalize the filter response for each time step by dividing it by its l2-norm. Then, we maximize the sparsity of the learnable filterbank's response by employing $l1$ penalty:

$$L_{indirect} = \frac{1}{N}\sum_{n=1}^{N} \left| \frac{O_{n,k}}{\sqrt{\sum_{k=1}^{K} O_{n,k}^2}} \right|_1 . \tag{4}$$

During the training, this objective is optimized by backpropagation with respect to the learnable parameterized filterbank $\mathbf{V}$. The term $O_{n,k}/\sqrt{\sum_{k=1}^{K} O_{n,k}^2}$ normalizes the vector of all filter responses at time step n to live on the unit-sphere. This suggests that this penalty is scale-invariant and insensitive to changing a overall gain of time-frequency features. Moreover, if a few responses tend to be significant, the others will decrease simultaneously. Thus, filters that model similar speaker-related frequency patterns with similar responses would get a high penalty. Since we also minimize $l1$ of the filter response, the sparsity of the learned time-frequency features is maximized.

Formally, the optimization goal of the training process is given by:

$$L = L_{sv} + \alpha(\beta L_{direct} + (1-\beta)L_{indirect}), \tag{5}$$

where both α and β are hyper-parameters used to adjust the importance of sparsity, and L_{sv} denotes speaker classification loss. For simplicity, we balance the contribution weights of those two sparsity constraints, i.e. β is fixed to 0.5.

3. Experiments and Results

3.1. Datasets

Experiments are conducted on the VoxCeleb [24, 25] and CNCeleb [26] datasets. For VoxCeleb dataset, we only use the VoxCeleb2-dev [25] for model training, while the VoxCeleb1 [24] is utilized to evaluate the performance. The VoxCeleb2 development dataset consists of over 2,000 hours of recordings from 5,994 English speakers under text-independent scenarios. CNCeleb is a text-independent Mandarin dataset and collects more than 130,000 utterances from 1,000 Chinese celebrities. It amounts to 274 hours in total. The training part involves 800 speakers, while the evaluation part contains 18,849 utterances from 200 speakers. To enrich the diversity, as described in [27], we augment the original CNCeleb dataset, as well as VoxCeleb2-dev using RIR and MUSAN datasets.

3.2. Metric

In this paper, both equal error rate (EER) and minimum detection cost function (minDCF) are employed to measure the performances of speaker verification systems. In consistence with [24], the target probability P_{tar} is set to 0.01, C_{fa} and C_{fr} share the equal weight of 1.0.

3.3. Implementation Details

The time-domain raw waveform input is sampled at 16K Hz. Conv-STFT [12] has 512 complex-valued kernels with a length of 400 (25ms) and a stride of 160 (10ms). Each parameterized kernel is implemented by the product of a Hamming window function and a corresponding complex-valued exponential STFT kernel. The dimension of magnitude output of each frame is $F = 257 = 512/2 + 1$. For a fair comparison with other published works, the total number of filters with different con-

Table 1: *Results on the Voxceleb1-O dataset using different parameterized filters. For a fair comparison, our experimental settings are consistent with the baselines, only except for the learnable filters.*

Model	Input Features	Parameterized Filter	EER (%)	minDCF
ResNet34 [12]	Magnitude spectrum	-	2.51	0.191
Res2Net [22]	Mel-FBank	-	1.45	0.147
ECAPA-TDNN	Magnitude spectrum	-	1.34	0.093
	Mel-FBank	-	1.17	0.082
	MFCC	-	1.30	0.086
ECAPA-TDNN	Magnitude spectrum	Vanilla filter (from scratch)	1.74	0.127
		Vanilla filter (Mel init.)	1.60	0.119
		Normalized filter	1.27	0.078
		Sparse filter	**1.03**	**0.073**

Table 2: *Results on the CNCeleb-E dataset and cross-language VoxCeleb1-O dataset. All the methods are trained on the augmented CNCeleb-dev dataset.*

Method	α	CNCeleb-E		VoxCeleb1-O	
		EER	minDCF	EER	minDCF
Mel-FBank	-	13.27	0.538	11.17	0.538
SF-L1	0.1	12.49	0.536	11.02	0.541
	0.3	12.37	0.531	10.79	0.524
	0.5	13.01	0.537	10.99	0.536
SF-L2	0.1	**12.25**	0.539	10.81	**0.520**
	0.3	12.66	0.537	**10.71**	0.533
	0.5	12.27	**0.527**	10.86	0.536
ICspk [12]		13.12	0.594	N/R	N/R
ResNet-DTCF [23]		14.84	0.596	N/R	N/R

straints is fixed to $K = 80$.

We use the ECAPA-TDNN [28] as the estimator of speaker identity and extractor of 192-dimensional speaker embeddings. In Table 1, we provide comparison of ECAPA-TDNN with older baselines. All the models are trained using AAM-softmax [29] with a margin of 0.2 and a scaling of 30. The Adam optimizer is employed with an initial learning rate of 0.001 to update the parameters of the speaker embedding extractor and learnable filters. The mini-batch size of 200 is chosen for all models training. For the back-end, we use cosine similarity. Our implementations (e.g. training strategies, augmentation, scoring) are consistent with [12, 27].

3.4. Analysis of different compression methods and parameterized filters

Before starting experiments with learnable filters, we compare older (ResNet34 and Res2Net) baselines, as well as ECAPA-TDNN system [28] using Mel-FBank and magnitude spectra. The comparison in the first two sections of Table 1 clearly shows the superiority of ECAPA-TDNN.

The SV performance of learnable filters is reported in the last section of Table 1. In addition, we also visualize the sorted frequency responses of those filters after the convergence in Fig 2. For experiments with unconstrained vanilla filters, it can be observed that they perform worse than the hand-crafted acoustic features (i.e. Mel-FBank, MFCC and STFT). It can be also observed (Fig 2 (b)) that the unconstrained filters contain a lot of sharp peaks and low-valued fluctuation, thus yielding poor performance. When using Mel-scale to initialize the parameters of filters, a slight improvement is achieved. The normalized filter clearly outperforms the unconstrained ones and the proposed LearnSF exhibits the best performance among all learnable filter-based systems.

3.5. Analysis on CNCeleb dataset

We experiment with various choices of hyper-parameter α and regularization of the filterbank. According to Eq. 3, two kinds of regularization are taken into consideration: SF-L1 denotes the $l1$-regularization on filter sparsity, and SF-L2 denotes $l2$-regularization. Also, the importance of sparsity constraint is analyzed through the SV performance. All the systems are trained on the CNCeleb dataset and evaluated on both CNCeleb-E and VoxCeleb1-O. The results are reported in Table 2.

It is observed that both SF-L1 and SF-L2 outperform the baseline SV model using the hand-crafted acoustic feature (such as Mel-FBank) on co-language and cross-language conditions. It suggests the superiority of LearnSF, which can provide an inductive bias to specific speech tasks. We also observe that SL-L1 performs slightly worse than SL-L2 in the current settings. This might be because the $l1$-regularization is not differentiable at zero, and hence, the optimization might fall into difficulty. Compared to other systems, the proposed approach achieves state-of-the-art performance on the CNCeleb dataset.

At the end of Table 2, we have included the results of two SOTA techniques evaluated in the CNCeleb-E dataset. ICSpk utilizes a set of modified complex exponential filters with learnable center frequency to model the time-domain waveform. It performs worse than our systems. This suggests it is difficult to model the raw waveform directly. Instead, our proposed method first transforms the time-domain signal into a frequency-domain and then tend to extract speaker-related pattern, which has the potential to guide more effective speaker feature selection. In addition, compared to ResNet-DTCF, the results of frequency-domain learned sparse filters are clearly superior.

4. Conclusion

In this paper, we propose a learnable sparse filterbank to directly model the magnitude spectrum. Two kinds of sparsity constraints are introduced to optimize the filterbank initialized on Mel-Scale: each filter is expected to activate a few frequency components and reduce others to zero while maintaining differentiation between filters. The SV experiments conducted on different datasets show the proposed system outperforms state-of-the-art systems by a significant margin.

5. References

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] Weicheng Cai, Jinkun Chen, and Ming Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[3] Junyi Peng, Rongzhi Gu, and Yuexian Zou, “Deep speaker embedding with long short term centroid learning for text-independent speaker verification,” *Proc. Interspeech 2020*, pp. 3246–3250, 2020.

[4] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, “Generalized end-to-end loss for speaker verification,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[5] Chunlei Zhang and Kazuhito Koishida, “End-to-end text-independent speaker verification with triplet loss on short utterances.,” in *Interspeech*, 2017, pp. 1487–1491.

[6] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, “Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification.,” *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[7] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu, “Improved rawnet with filter-wise rescaling for text-independent speaker verification using raw waveforms,” in *Proc. Interspeech*, 2020, pp. 1496–1500.

[8] Weiwei Lin and Man-Wai Mak, “Wav2spk: A simple dnn architecture for learning speaker embeddings from waveforms,” *Proc. Interspeech 2020*, pp. 3211–3215, 2020.

[9] Mirco Ravanelli and Yoshua Bengio, “Speaker recognition from raw waveform with sincnet,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[10] Erfan Loweimi, Peter Bell, and Steve Renals, “On learning interpretable cnns with parametric modulated kernel-based filters.,” in *INTERSPEECH*, 2019, pp. 3480–3484.

[11] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux, “Learning filterbanks from raw speech for phone recognition,” in *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2018, pp. 5509–5513.

[12] Junyi Peng, Xiaoyang Qu, Jianzong Wang, Rongzhi Gu, Jing Xiao, Lukáš Burget, and Jan Černockỳ, “Icspk: Interpretable complex speaker embedding extractor from raw waveform,” *Proc. Interspeech 2021*, pp. 511–515, 2021.

[13] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, “Leaf: A learnable frontend for audio classification,” *arXiv preprint arXiv:2101.08596*, 2021.

[14] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F Lyon, and Rif A Saurous, “Trainable frontend for robust and far-field keyword spotting,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.

[15] Kin Wai Cheuk, Hans Anderson, Kat Agres, and Dorien Herremans, “nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks,” *IEEE Access*, vol. 8, pp. 161981–162003, 2020.

[16] Hong Yu, Zheng-Hua Tan, Yiming Zhang, Zhanyu Ma, and Jun Guo, “Dnn filter bank cepstral coefficients for spoofing detection,” *Ieee Access*, vol. 5, pp. 4779–4787, 2017.

[17] Jiquan Ngiam, Zhenghao Chen, Sonia Bhaskar, Pang Koh, and Andrew Ng, “Sparse filtering,” *Advances in neural information processing systems*, vol. 24, 2011.

[18] Junyi Peng, Xiaoyang Qu, Rongzhi Gu, Jianzong Wang, Jing Xiao, Lukáš Burget, and Jan Černocký, “Effective Phase Encoding for End-To-End Speaker Verification,” in *Proc. Interspeech 2021*, 2021, pp. 2366–2370.

[19] Erfan Loweimi, *Robust Phase-based Speech Signal Processing From Source-Filter Separation to Model-Based Robust ASR*, Ph.D. thesis, University of Sheffield, 2018.

[20] Quchen Fu, Zhongwei Teng, Jules White, Maria Powell, and Douglas C Schmidt, “Fastaudio: A learnable audio front-end for spoof speech detection,” *arXiv preprint arXiv:2109.02774*, 2021.

[21] Tara N Sainath, Brian Kingsbury, Abdel-rahman Mohamed, and Bhuvana Ramabhadran, “Learning filter banks within a deep neural network framework,” in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 297–302.

[22] Tianyan Zhou, Yong Zhao, and Jian Wu, “Resnext and res2net structures for speaker verification,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.

[23] Li Zhang, Qing Wang, and Lei Xie, “Duality Temporal-channel-frequency Attention Enhanced Speaker Representation Learning,” *arXiv preprint arXiv:2110.06565*, 2021.

[24] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: A large-scale speaker identification dataset,” *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[25] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[26] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[27] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, “In defence of metric learning for speaker recognition,” in *Interspeech*, 2020.

[28] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” in *Interspeech2020*, 2020, pp. 1–5.

[29] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.