INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

ICSpk: Interpretable Complex Speaker Embedding Extractor from Raw Waveform

Junyi Peng[1,2,†], *Xiaoyang Qu*[1], *Jianzong Wang*[1*], *Rongzhi Gu*[3] , *Jing Xiao*[1], *Lukáš Burget*[2], *Jan "Honza" Černocký*[2]

[1]Ping An Technology (Shenzhen) Co., Ltd., China
[2]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[3]Peking University, Shenzhen Graduate School, China

pengjy@fit.vutbr.cz, {quxiaoyang343,wangjianzong347,xiaojing661}@pingan.com.cn

Abstract

Recently, extracting speaker embedding directly from raw waveform has drawn increasing attention in the field of speaker verification. Parametric real-valued filters in the first convolutional layer are learned to transform the waveform into time-frequency representations. However, these methods only focus on the magnitude spectrum and the poor interpretability of the learned filters limits the performance. In this paper, we propose a complex speaker embedding extractor, named ICSpk, with higher interpretability and fewer parameters. Specifically, at first, to quantify the speaker-related frequency response of waveform, we modify the original short-term Fourier transform filters into a family of complex exponential filters, named interpretable complex (IC) filters. Each IC filter is confined by a complex exponential filter parameterized by frequency. Then, a deep complex-valued speaker embedding extractor is designed to operate on the complex-valued output of IC filters. The proposed ICSpk is evaluated on VoxCeleb and CNCeleb databases. Experimental results demonstrate the IC filters-based system exhibits a significant improvement over the complex spectrogram based systems. Furthermore, the proposed ICSpk outperforms existing raw waveform based systems by a large margin.

Index Terms: end-to-end speaker verification, raw waveform, complex neural networks, interpretable complex filters

1. Introduction

Speaker verification (SV) is a process to verify whether an unknown utterance belongs to its claimed identity. According to the application scenario, SV can be categorized to the text-dependent speaker verification (TD-SV) and text-independent speaker verification (TI-SV) [1]. Since TI-SV has no constraint of transcripts, compared to TD-SV, it has greater potential in applications. In this paper, we focus on TI-SV.

As an efficient statistical model, i-vector+PLDA has achieved great success in TD-SV task [2]. Recently, as deep learning shows its remarkable success in speech modeling, more researchers focus on building deep structures [3, 4] or investigating effective objective functions [5, 6] to extract discriminant speaker representations. Most of these approaches employ hand-crafted acoustic features, such as log Mel filterbank (Fbank) and Mel frequency cepstral coefficients (MFCC). One potential shortcoming of these methods is that, such representations (e.g., FBank, MFCC) are data-independent because of predefined and fixed feature extraction parameters, which indicates they cannot be trained forwards specific speech-related task. Additionally, these features could lose useful acoustic information during the nonlinear transform, which may lead to a performance bottleneck of SV systems.

To mitigate this problem, one natural approach is to learn the acoustic features automatically as a part of neural network model. For better incorporating with SV task, the network should model the vocal tract related characteristics directly from the raw waveform with a set of learnable filters in the first layer [7]. Following this pipeline, in [8], a carefully designed convolution neural network (CNN) is used to directly model the raw waveform. It exhibits good performance on VoxCeleb1 dataset [9]. In [10], a feature encoder consisting of convolutional layers with large stride and kernel size is leveraged to deal with the raw waveform, which obtains a comparable result to start-of-the-art systems. Although the learned filters outperform the hand-engineered filters, the training process is unstable and the interpretability in time and frequency about those structures remains shallow.

To enhance the interpretability, the original convolutional filters have been modified to learn well-explored acoustic features via incorporating prior signal processing knowledge. In [11], a novel convolution layer composed by parameterized band-pass filters, named SincNet, is designed to obtain speaker embedding. Compared to traditional CNN, SincNet takes the advantages of a parametric model: higher interpretability and fewer parameters [12]. [13] utilizes power Gabor filter to craft learnable spectrograms for audio classification task. Each Gabor filter has only two parameters, i.e., center frequency and inverse bandwidth. The system reaches state-of-the-art performance on the AudioSet benchmark. Nevertheless, these methods only consider the magnitude spectrum, while the importance of phase in the complex-valued spectrogram is neglected.

In this paper, we build upon our previous research [14] and design a novel complex-valued deep neural network to extract speaker embeddings with high interpretability, named *interpretable complex speaker embedding extractor* (ICSpk). The basic idea is to learn interpretable complex (IC) filters based on the well-defined short-term Fourier transform (STFT). According to the definition of STFT, the waveform is decomposed on a set of complex exponential bases, the frequency responses of which are evenly distributed. However, as [15] points out, the speaker information is located mostly in the low-frequency region. Therefore, to distinguish speakers more precisely in low-frequency region, we propose to learn a new frequency response distribution for complex filters in data-driven fashion. Specifically, the complex exponential bases are implemented

[†] work done during internship at Ping An Technology
[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2016

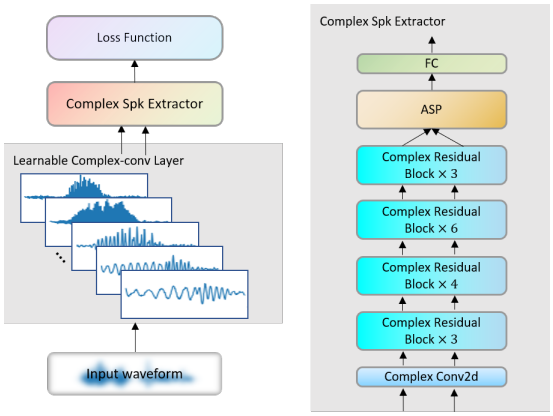


Figure 1: *The illustration of interpretable complex speaker embedding extractor, which consists of interpretable complex-convolutional layer, complex speaker encoder. The details of complex residual block are shown in Figure 2.*

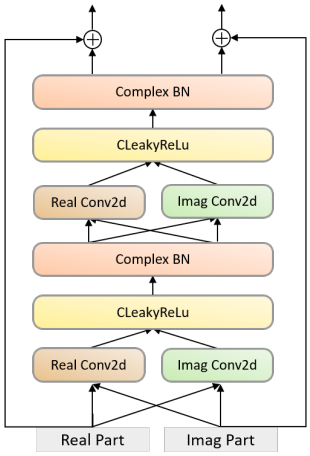


Figure 2: *Details of the complex residual block. "CLeakyReLu" is the complex leaky ReLu. "Complex BN" means complex batch normalization.*

as interpretable complex (IC) filters of a convolutional layer. The frequency component of each filter is set as a learnable parameter, which is initialized based on the original STFT definition. These IC filters directly operate on the raw waveform and produce a complex-valued time-frequency representation that is optimized for SV task. With such representation, a dedicated complex-valued convolutional neural network, which combines the advantage of both complex neural networks and residual connections, is designed to further extract the speaker information in complex domain. Extensive experiments are conducted on two large-scale TI-SV datasets: VoxCeleb [9, 16] and CNCeleb [17]. Results show that the proposed ICSpk consistently outperforms feedforward methods based on non-adaptive traditional features, as well as the state-of-the-art raw waveform based methods.

The rest of paper is organized as follows: Section 2 gives a brief introduction to the SincNet. Section 3 describes the proposed ICSpk in detail. Experimental setup including database description, training paradigm, and result analysis are described in Section 4 and 5. Section 6 concludes the paper.

2. Related Work

Standard CNNs operate on the raw waveform by performing time-domain convolutions between the input waveform and a set of certain finite impulse response filters [18]. The first layer of **SincNet** employs a set of band pass filters implemented by two sinc functions, resulting in an ideal band-pass filter. The impulse response $g(t)$ of the band-pass filter is:

$$g(t) = 2f_2 sinc(2\pi f_2 t) - 2f_1 sinc(2\pi f_1 t) \tag{1}$$

where the *sinc* function is defined as $sinc(x) = \frac{sin(x)}{x}$, f_1 and f_2 are learnable parameters, which denote the low and high cutoff frequency respectively. Note that $f_2 > f_1$. Eq.1 in the frequency domain gives:

$$G(f) = rect(\frac{f}{2f_2}) - rect(\frac{f}{2f_1}) \tag{2}$$

where $rect(\cdot)$ is the rectangular function. $G(\cdot)$ is the frequency response. However, SincNet only considers the real-valued magnitude part of complex-value spectrogram.

3. Interpretable Complex Speaker Embedding Extractor

In this section, we will describe the proposed complex-valued system, which directly models the raw speech signal, as shown in Figure 1. At first, we introduce the proposed interpretable complex filter. Then, we describe the complex speaker embedding extractor.

3.1. Interpretable complex-convolution filter

As mentioned in [11], conventional CNN filters are not as effective in capturing fundamental acoustic features from raw waveform as expected. This is due to the lack of constraint to the learnable parameters. Intuitively, the frequency response of the widely-used STFT is evenly distributed. However, most of speaker-related information lays in low-frequency regions [19]. To take advantage of this characteristic, we redesign the STFT kernel to emphasise the information in low frequency. Specifically, a family of IC filters incorporating prior knowledge from STFT are employed to directly deal with raw waveform and produce a complex-valued time-frequency representation. The center frequency of each IC filter is set as the learnable parameter. Mathematically, an IC filter with learnable real-valued parameter k is defined as follow:

$$X[n] = \sum_{m=0}^{N-1} x[m]\omega[n-m]e^{-ikn} \tag{3}$$

where $x[n]$ is the input raw waveform, $w[n]$ denotes the window function with the length N. In this paper, we use the well-known Hanning window. $X[n]$ is the complex-valued time-frequency representation of $x[n]$ processed by the IC filters. The IC filters can be split into real F_{real} and imaginary F_{imag} part, respectively:

$$\begin{aligned} F_{real}[n,k] &= \omega[n]\cos[kn] \\ F_{imag}[n,k] &= -\omega[n]\sin[kn] \end{aligned} \tag{4}$$

where the filter length of F_{real} and F_{imag} is decided by the window function $\omega[n]$ with the length N, and $n \in [0, N-1]$

denotes the time index in window w. The stride of IC filters makes reference to the hop size in STFT.

3.2. Complex speaker embedding extractor

To consistently work with complex-valued features derived from IC filters, the most straightforward approach is to concatenate the real and imaginary parts on the channel axis and then feed them to the standard (real-valued) speaker embedding extractor. During this process, the complex multiplication rule is ignored. The neural network may not able to learn the correlation between the real and imaginary parts. This could limit the performance of the speaker embedding extractor.

To enhance the interaction between the real and imaginary parts contained in complex-valued features, we redesign the convolutional layers, activation functions and normalization within the residual blocks of speaker embedding extractor to handle the complex domain operations. As shown in Figure 2, the complex convolution kernel $\boldsymbol{W}$ is defined as $\boldsymbol{W} = \boldsymbol{A} + i\boldsymbol{B}$, where the real and imaginary parts of a complex kernel are implemented by the real-valued matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. The complex operation on complex input matrix $\boldsymbol{H} = \boldsymbol{X} + i\boldsymbol{Y}$ is defined as:

$$\begin{aligned} \boldsymbol{W} \otimes \boldsymbol{H} &= (\boldsymbol{A} \otimes \boldsymbol{X} - \boldsymbol{B} \otimes \boldsymbol{Y}) + i(\boldsymbol{A} \otimes \boldsymbol{X} + \boldsymbol{B} \otimes \boldsymbol{Y}) \\ &= \boldsymbol{P} + i\boldsymbol{Q} \end{aligned} \tag{5}$$

where $\otimes$ denotes the real-valued convolution operation, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are real and imaginary parts of the complex convolution product, respectively. The complex residual block (CRS) [20] consists of two repeats of one complex convolutional layers followed by complex batch normalization (ComplexBN) [21] and complex leaky rectified linear unit (CLeakyReLu). Additionally, a skip connection is added between the block input and output.

The complex speaker embedding extractor, an advanced ResNet34 structured model, employs 1 complex convolution layer and 4 stacks of CRSs to generate the frame-level features. The number of channels in these CRSs is set as 8, 16, 32, and 64, respectively. The size of complex convolution kernels in CRS is always set to 3×3. The frame-level features are aggregated into an utterance-level representation through an attentive statistics pooling (ASP) layer [22]. Then a feedforward layer is utilized to obtain speaker embedding.

3.3. Properties

In summary, the proposed ICSpk has some interesting properties:

Fewer Paramters. The IC filters have fewer parameters than sinc-conv filters. Specifically, each filter of sinc-conv has two learnable parameters (i.e. the low and high cutoff frequencies), while the IC filter has only one learnable parameter. Moreover, with increasing filter length, the number of parameters of conventional 1-dimensional CNN kernel filter grows proportionally, while IC filter has still only one parameter.

IC filter has the unchanged parameter.

Higher Interpretability. The IC filter is derived from the STFT filter. The response of each filter corresponds to the task-related region, which has a clear physical meaning.

4. Experiments

For all experiments, we use the same experimental setup to perform fair comparison. The data processing, training and testing strategies for all presented experiments are the same.

4.1. Dataset

Experiments are conducted on the VoxCeleb [9, 16] and CNCeleb [17] datasets. For VoxCeleb dataset, we only use the VoxCeleb2 [16] for model training, while the VoxCeleb1 [9] is utilized to evaluate various protocols. The VoxCeleb2 development dataset is collected from YouTube and contains over 2,000 hours of recordings from 5,994 English speakers under text-independent scenarios. The VoxCeleb1 dataset contains 1,251 speakers, and is used to construct verification trials.

CNCeleb is a large-scale text-independent dataset and contains over 130,000 utterances from 1,000 Chinese celebrities from BiliBili. It covers 11 genres and the total duration of speech is 274 hours. The training part contains 800 speakers, while the evaluation part contains 18,849 utterances from 200 speakers. To increase the diversity, we augment the original CNCeleb dataset using RIR and MUSAN datasets.

4.2. Model training

The duration of input raw waveform ranges from 200 to 400 ms. The mini-batch size is 120. The dimension of speaker embedding is set as 512. Following [23], angular prototypical (AP) loss is selected as objective function of the ICSpk. Moreover, we choose Adam as the optimizer with an initial learning rate of 0.001. L2 regularization is applied to prevent overfitting with the rate of $5e^{-5}$. To speed up the training process, the learning rate is decreased by 10% every 2 epochs. The parameters of IC filters are initialized using original STFT configuration (i.e. k_j is initialized with $j2\pi/N$, where k_j is the learnable parameter of the j-th IC filter, N is the length of IC filter), and the rest of neural network is initialized with the default initialization in PyTorch. The models are trained on 8 NVIDIA Tesla V100 GPUs for 50 epochs. It is noted that to keep the same parameters with the original ResNet34 (1.9 M), we reduce all the convolution channels by half for fair comparison. For VoxCeleb SV, the kernel size (window length) of the complex filters is set to 400, the stride (hop size) is 160, the total number of complex filters is 512.

5. Results

5.1. Metric

Equal error rate (EER) and minimum detection cost function (minDCF) are used to measure the speaker verification system performance. The target probability P_{tar} is 0.01, C_{fa} and C_{fr} have the same weight of 1.0, which is a standard setting [9].

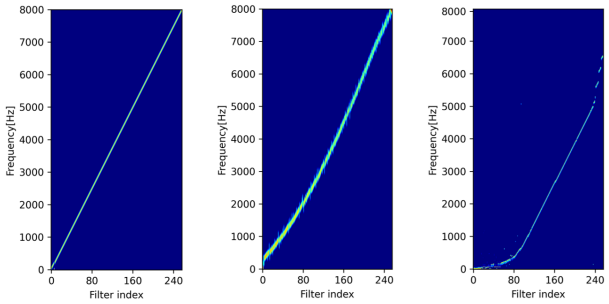
5.2. Comparison with state-of-the-art systems

To demonstrate the effectiveness of proposed ICSpk, we compare it with other state-of-the-art systems using raw waveform as input. Table 1 reports the results of our proposed ICspk with ResNet34 using different types of input acoustic features, including magnitude, concatenation of real and imaginary parts of complex spectrogram and raw waveform. We also list the state-of-the-art SV systems using raw waveform (RawNet2, Wav2spk, raw-x-vector) for comparison.

Firstly, using real-valued ResNet34 as front-end model, "Real+Imag" based system outperforms "Magnitude" based system (i.e. 2.34% v.s. 2.51%), implying that the phase part is also embedded with speaker related information that has been neglected. Learning band-pass filters (sinc-conv) outperforms the magnitude based system, while exhibiting slight worse performance than "Real+Imag" based system. Replacing the real-

Table 1: *Results for speaker verification on the Voxceleb1 dataset and extended VoxCeleb1-E and VoxCeleb-H test sets. N/R : Not report results. ℂResNet34: complex ResNet34. AP: Angular Prototypical.*

Front-end Model	Params	Input Feature	Loss	VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H	
				EER	minDCF	EER	minDCF	EER	minDCF
RawNet2 [24]	N/R	Raw waveform	Softmax	2.48	N/R	2.57	N/R	4.89	N/R
Wav2spk [10]	N/R	Raw waveform	AM-softmax	1.95	0.203	N/R	N/R	N/R	N/R
raw-x-vector [25]	4.2M	Raw waveform	AM-softmax	2.56	0.25	2.41	0.25	3.99	0.36
ResNet34	1.9M	Magnitude	AP	2.51	0.191	2.55	0.194	4.89	0.323
ResNet34	1.9M	Real + Imag	AP	2.34	0.178	2.34	0.171	4.40	0.278
ResNet34	1.9M	Sinc-conv	AP	2.36	0.183	2.55	0.188	5.01	0.317
ℂResNet34	1.9M	Real + Imag	AP	2.02	0.137	2.09	0.151	3.94	0.254
ICspk	1.9M	Raw waveform	AP	**1.92**	**0.137**	**1.94**	**0.141**	**3.78**	**0.237**



(a) original STFT (b) Sinc filters (c) IC filters

Figure 3: *Frequency response of three different types of filters. (a) Vanilla STFT filters. (b) Learned Sinc-convolution filters. (c) Learned IC filters. Filters are sorted by their corresponding peak values in the frequency domain and the number of filters is fixed to N = 257.*

valued speaker embedding extractor (ResNet34) with the proposed ℂResNet34, 14% relative improvement (i.e. 2.02% v.s. 2.34%) is achieved with the same number of parameters. Finally, by integrating the proposed IC filters and complex network structure, the proposed ICSpk outperforms all other raw waveform based systems and achieves state-of-the-art performance on VoxCeleb1 dataset.

In order to further analyze how and why our proposed method is effective, we visualize the frequency response of original STFT filters, Sinc-conv filters and proposed complex-conv filters by sorting their corresponding peak values in Figure 3. The frequency response of original STFT is evenly distributed, as defined. For the learnable filters, the learned sinc-conv filters share the similar property with the learned IC filters. The majority of filters are tuned to lower frequencies, conforming that the low frequency features play a critical role in distinguishing speaker. In detail, since there are two learnable parameters, i.e., upper and lower bounds, the responses of sinc-conv filters tend to be flocculent. The complex-conv filter including one learnable parameter provides a higher resolution in low frequency.

5.3. The effect of window length

A potential pitfall of the STFT is that it has a fixed resolution determined by a predefined window length: a wide window gives a good frequency resolution but poor time resolution, and vice versa. Since the complex-convolution layer convolves the raw waveform with a set of parametrized STFT bases, we report our experimental results on how the window length and learnable parameters affect the performance of the SV systems. Three window length settings (i.e. 128, 256, 512) are compared. We use the ResNet34 to extract the speaker embeddings. Table 2 presents the results evaluated on CNCeleb(E). The bold font denotes the best result when the loss function is fixed.

Table 2: *Effect of window length on CNCeleb.Eval test dataset. ResNet34 is employed as feature extractor. L denotes filter length of IC filters*

Model	L	Learnable	EER	minDCF
i-vector[17]	-	-	14.24	N/R
x-vector[17]	-	-	14.78	N/R
ResNet34[26]	-	-	16.51	N/R
ResNet34	128	✓	15.55	0.636
		-	14.31	0.627
	256	✓	**13.12**	0.611
		-	14.41	0.626
	512	✓	13.31	**0.594**
		-	13.85	0.621

As shown in Table 2, the wide window is able to surpass the performance of narrow one. This suggests that the fine structure of frequency is beneficial to the speaker representation. In terms of the learnable parameters, almost all results show that using a set of learnable complex filters gives a better result. This indicates that having a flexible bias to the SV, the model has a potential to produce more discriminative speaker embeddings. Moreover, the proposed system outperforms the previous state-of-the-art systems by large margin.

6. Conclusion

In this paper, we propose interpretable complex filters derived from the STFT kernel to directly model the raw waveform. The IC filter in the first convolution layer is implemented using a complex exponential whose frequency is learnable. Besides, a complex speaker embedding extractor is proposed to deal with the complex output of IC filters. The conducted SV experiments show the proposed system outperforms state-of-the-art systems while operating on the raw waveform by a large margin.

7. Acknowledgements

This work was 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".

8. References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[3] W. Cai, J. Chen, and M. 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.

[4] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[5] J. Peng, R. Gu, and Y. Zou, "Deep speaker embedding with long short term centroid learning for text-independent speaker verification," *Proc. Interspeech 2020*, pp. 3246–3250, 2020.

[6] L. Wan, Q. Wang, A. Papir, and I. L. 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.

[7] D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "A deep 2d convolutional network for waveform-based speech recognition," *Proc. Interspeech 2020*, pp. 1654–1658, 2020.

[8] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification." *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[10] W. Lin and M.-W. Mak, "Wav2spk: A simple dnn architecture for learning speaker embeddings from waveforms," *Proc. Interspeech 2020*, pp. 3211–3215, 2020.

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

[12] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable cnns with parametric modulated kernel-based filters." in *INTERSPEECH*, 2019, pp. 3480–3484.

[13] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[14] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[15] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[17] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. 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.

[18] L. Rabiner and R. Schafer, *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[19] H. Muckenhirn, M. M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4884–4888.

[20] C. Trabelsi, O. Bilaniuk, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *CoRR*, 2017.

[21] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[22] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018, pp. 2252–2256.

[23] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2977–2981.

[24] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech 2020*, pp. 1496–1500, 2020.

[25] G. Zhu, F. Jiang, and Z. Duan, "Raw-x-vector: Multi-scale time domain speaker embedding network," *arXiv preprint arXiv:2010.12951*, 2020.

[26] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *arXiv preprint arXiv:2012.12468*, 2020.