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

Effective Phase Encoding for End-to-end Speaker Verification

Junyi Peng[1,2,†], *Xiaoyang Qu*[1], *Rongzhi Gu*[3], *Jianzong Wang*[1*], *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

The widely used magnitude spectrum based features have shown their superiority in the field of speech processing. In contrast, the importance of phase spectrum is always ignored. This is because the patterns hidden in phase cannot be intuitively modelled and interpreted, due to phase wrapping phenomenon. In this paper, we explore novel phase spectrum based features, named Learnable Group Delay (LearnGD), to capture useful information in speech signals. Specifically, firstly, the negative of the spectral derivative of the phase spectrum, called group delay (GD), is used to unwrap the phase. Then, to suppress the spiky nature of GD, which is caused by its roots close to the unit circle in the Z domain, a carefully designed light convolutional smoothing layer is employed to reconstruct the GD. Finally, an exponential hyper-parameter is introduced to reconstruct GD features to restore the spectrum range and generate LearnGD features. For performance evaluation, speaker verification experiments are conducted on the VoxCeleb2 corpus. Compared to the traditional acoustic feature derived from the magnitude spectrum, the proposed phase-based features reach a 27.8% relative improvement in terms of EER. Furthermore, experimental results on TIMIT phoneme recognition task also demonstrate the effectiveness of our proposed phase-based features.

Index Terms: end-to-end speaker verification, phase information, group delay, on-the-fly

1. Introduction

Speaker verification (SV) is the process of verifying whether an unknown speech belongs to a specific target speaker. According to the restriction of the content, speaker verification can be categorized into text-dependent speaker verification (TD-SV) and text-independent speaker verification (TI-SV) [1]. This paper focuses on TI-SV.

For many years, the combination of i-vector and Probabilistic Linear Discriminant Analysis (PLDA) has been the dominant approach [2]. Recently, with the advance in deep learning, more attention has been paid to discriminative speaker embedding learning in SV task. In order to enhance the discrimination of the speaker embedding, researchers investigate neural network structures and loss functions (e.g. triplet loss[3], angular softmax loss [4], affinity loss[5]).

Most of these state-of-the-art SV systems use spectral features derived from short-term Fourier transform (STFT) power spectrum, such as filterbank, Mel Frequency Cepstrum Coefficient (MFCC). It has been shown that the phase information in speech influences the intelligibility [6]. In addition, a time domain speech signal can be recovered uniquely only if both magnitude and phase spectrum are known. This indicates the phase spectrum has potential to encode some acoustic information [7, 8], which may benefit to speech-related tasks. However, extracting useful information from phase spectrum is not straightforward due to the phase wrapping phenomenon. Compared to the magnitude spectrum, phase spectrum has an intractable and noise-like shape, which is hard to interpret and model directly [9].

To solve this problem, researchers turned to other phase-related representations, which contain the majority of phase information, while being more tractable. In [10], the phase is mapped into a polar coordinate on a unit circle. This modified phase feature is more robust than the original phase features for various speaker systems on the NTT dataset. It has been shown in [11], that the modified phase-based system provides a good performance in NIST SRE 2010 when fused with MFCC. In [12], the instantaneous frequency cosine coefficients (IFCC) features, which are extracted from the analytic phase of speech, are utilized to capture the subtle acoustic variations in live and replayed speech. The IFCC-based system leads to better results than the MFCC-based system in ASVspoof 2017 corpus.

To further investigate the information contained in phase spectrum, in the last few years, the negative derivative of the phase spectrum, named group delay (GD), and its varieties are raising attention to representing meaningful properties of speech from phase spectrum [13, 14, 15]. The series of GD features have the characteristics of high frequency resolution and low frequency leakage at the same time. However, due to influence of window function and noise in practice, the GD features may be spiky when their zeros are close to unit circle in z-domain, which will fuzz the fundamental frequency and other useful acoustic information. In [13], a modified GD (MODGD) features are proposed to solve this spikiness issue through cepstral smoothing. A potential shortcoming of these features is that the parameters are fixed congenitally and not learnable using the training data. The feature extraction is independent of model training and does not provide any proper bias to the specific speech task. In addition, cepstral smoothing algorithm that implemented through a median filter, DCT and inverse DCT is time consuming. The range of MODGD is still uncontrollable.

In this paper, we propose a novel phase-based feature, called learnable group delay (LearnGD), to solve these problems. Firstly, we unwrap the original phase spectrum using group delay function. Then, since the GD is undefined when the roots of transfer function get close to unit circle, we design a light convolutional smoothing layer to efficiently filter

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

http://dx.doi.org/10.21437/Interspeech.2021-2025

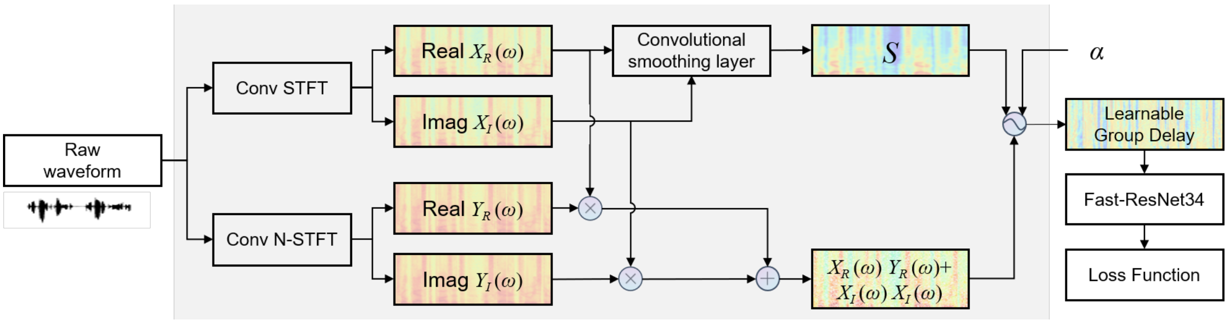


Figure 1: *The illustration of speaker verification framework with learnable group delay features*

out the excitation component in GD and introduce an exponential hyper-parameter to regulate the dynamic range. We expect the LearnGD to be capable of pulling the roots of speech away from the unit circle in Z domain to avoid spikes which cause distortion of formats, while having a more flexible inductive bias that is able to enhance the task-related vocal tract representation and weaken the irrelevant part in the GD domain. Moreover, in order to combine the feature extraction and model training into one cascaded pipeline, motivated by [16, 17], we implement the STFT operation by 1-dimensional convolution layer. In this way, the whole calculation process of LearnGD feature is done on the GPU devices, which allows the end-to-end neural network training (e.g. directly take the raw waveform as input) and speeds up the forward processing. Extensive experiments are conducted on the VoxCeleb1&2 SV task and TIMIT phoneme recognition task. Results show that our proposed LearnGD outperforms widely used acoustic features.

2. Preliminaries

2.1. Short-Time Fourier Transform

Given a time domain speech signal $x(n)$, its short-time Fourier transform (STFT) $X(n,\omega)$ after applying window function $w(n)$ with the length N, is defined as follows:

$$\begin{aligned} X(n,\omega) &= \sum x(m)w(n-m)e^{-j\omega m} \\ &= |X(n,\omega)|e^{j\theta(n,w)} \end{aligned} \tag{1}$$

where $|X(n,\omega)|$ is the magnitude spectrum and $\theta(n,w)$ is denoted as phase spectrum.

2.2. Group Delay

Group delay is defined as the negative derivative of the phase spectrum $\theta(w)$. It can be expressed as:

$$\begin{aligned} \tau(\omega) &= -\frac{d(\theta(\omega))}{d\omega} \\ &= -\mathrm{Im}\frac{\log(X(\omega))}{d\omega} \\ &= \frac{X_R(\omega)Y_R(\omega)+X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \end{aligned} \tag{2}$$

where $X_n(\omega)$ and $Y_n(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$. The subscripts I and R mean the real and imaginary parts, respectively. For a time delay system, the group delay is related to the time delay value for all the frequency components.

3. Learnable Group Delay Features

The framework of our proposed LearnGD feature based SV system is illustrated in Fig.1. At first, the input waveform $x(n)$ is passed to two 1-dimensional convolution layers that realize the STFT operation of $x(n)$ and $nx(n)$ separately, and then the corresponding complex spectrograms are generated. Secondly, a light convolutional smoothing layer is designed to smooth the power spectrogram of $x(n)$. Then, along with the hyper-parameter α, we compute the LearnGD feature, which is finally fed to a deep neural network to predict the speaker label. The whole operation is implemented on GPU, which enables the on-the-fly feature extraction and model training at the same time.

3.1. Convolution STFT Layer

In this paper, the STFT operation is implemented on-the-fly via expressing the vector multiplication as 1-dimensional linear convolution operation. The size of convolutional kernel is equal to the length of window function. We set the stride of convolution according to the hop size of STFT operation. The size and number of kernels can be customized according to different STFT configurations. The window function $w(n)$ can be implemented by multiplying $w(n)$ element-wise with the convolution kernels during the processing.

3.2. Learnable Group Delay Features

As mentioned in introduction, Group delay features suffer from unexpected spikes in spectrum that limit their performance in applications.

When the root (pole or zero) of Z-transformation of the speech signal (i.e., the denominator of Eq. 2) approaches the unit circle in Z domain, the GD value of the corresponding frequency component will become spiky and unstable. Furthermore, since the group delays of roots are added together, the entire group delay will become sharp even when one of the group delays is spiky. Thus, fundamental frequency and the fine structure may be obscured, ultimately reducing the effectiveness of obtained feature.

These spikes brought by zeros cannot be eliminated by normal linear smoothing functions. Hence, in [13], MODGD is proposed to address this spikiness issue through cepstral smoothing. However, the cepstral smoothing process, which includes a median filter, DCT and inverse DCT operation, is time-consuming. Also, the dynamic range of MODGD is still uncontrollable. To solve these problems, firstly, we design a light smoothing layer composed by a regular convolution layer with a normalized kernel K, which acts as limited-region attention module to smooth the power spectrum. The smoothed power spectrum can be obtained as follows:

$$S(n,\omega) = \textit{softmax}\,(K(n,\omega)) \otimes |X(n,\omega)|^2$$
$$S(n,f) = \sum_{\hat{f}=-F}^{F} \sum_{\hat{n}=-L}^{L} \textit{softmax}(K(n-\hat{n}, f-\hat{f})) \left|X(\hat{n},\hat{f})\right|^2 \tag{3}$$

where $f = \frac{N\omega}{2\pi}$, N is the window length, and $|X(n,f)|^2$ is the power spectrum, $\otimes$ is the convolution operator, $\textit{softmax}(\cdot)$ is the softmax activation function. The receptive field of the convolutional kernel K is $2L \times 2F$, where $2L$ and $2F$ denote the smoothing range along frame and frequency axis, respectively. Such a simple conventional kernel based normalization can effectively alleviate the spikiness problem by reducing the weight of abnormal regions, while generating more discriminative phase-based features.

To further restrict the dynamic range of group delay function, we introduce an exponential hyper-parameter $\alpha \in (0,1]$ for computing LearnGD feature by amplifying the low-value region and compressing the high-value region. Thus, the fine structure and spectral envelope of LearnGD can be emphasized at the same time. The learnable group delay can be finally defined as:

$$\tau_{\textit{LearnGD}} = |\frac{X_R(n,f)Y_R(n,f) + X_I(n,f)Y_I(n,f)}{S(n,f)}|^{\alpha} \tag{4}$$

4. Experiments and Discussion

In order to fairly compare the experimental results, we decided to make our experimental settings consistent with the baseline from [18], only except for the input feature. Thus, we utilize the same network structure, data processing procedure, loss function, training and testing strategies in our experiments.

4.1. Datasets

The SV performance is evaluated on the VoxCeleb corpus [19, 20], which is a widely used large-scale text-independent speaker verification dataset. The entire dataset involves two parts: VoxCeleb1 and VoxCeleb2. The utterances are collected from YouTube videos, where the celebrities belong to different races and have a wide range of accents. The training set is derived from the development set of VoxCeleb2. The TIMIT dataset contains 6300 utterances (5.4 hours), consisting of 10 utterances spoken by each of 630 speakers from 8 different regions.

4.2. Implementation details

Network structure: For speaker verification, we use Fast-ResNet34 [18] as the trunk architecture, which is a modified version of the original Thin-ResNet [4, 21]. To be specific, the input dimensions and strides are redesigned to reduce the computation cost. The output of the last hidden layer is extracted as the speaker embedding. No LDA nor PLDA is used. For TIMIT phoneme recognition task, our implementation is based on Pytorch-Kaldi standard recipe [22, 23]. The ASR feature extractor starts with 4 1D-CNN layers with 128, 60, 60, 60 convolution kernels of the size 129, 5, 5 and 3, respectively. Five additional feedforward layers followed with a softmax layer are used to predict the probability over HMM states.

Training: Our system is optimized by Adam, where the initial learning rate is 0.001, reduced by 4% every epoch. The Angular Prototypical (AP) [18] is utilized as loss function. The min-batch size is set to 160. L2-regularization is added to prevent overfitting during the training. For phoneme recognition, the CNN models are fed with 200ms waveforms with 10ms frame shift. A dropout rate of 0.15 is set for all layers expect the softmax layer. The mini-batch size is set to 128 for 23 epochs. The models are optimized by RMSProp with learning rate of 0.001. In this paper, we fix the F to 1 and denote the $2L$ as the smooth length. The parameters of the convolutional smoothing layer are initialized with the same value $\frac{1}{4L}$.

Metric: Equal error rate (EER) and minimum detection cost function (minDCF) are used to measure the speaker verification system performance. We use the same parameters as [19], where the target probability P_{tar} is 0.01, C_{fa} and C_{fr} have the same weight of 1.0. In ASR task, phone error rate (PER) is the most common evaluation measures.

4.3. Analysis of hyper-parameters

Table 1: *SV performances using different hyper-parameters settings on VoxCeleb1-O with feature extractor Fast-ResNet34.*

Length($2L$)	α	**EER**(%)	**minDCF**
80	0.2	2.76	0.215
	0.4	2.09	0.155
	0.6	2.17	0.159
100	0.2	2.29	0.167
	0.4	2.08	0.162
	0.6	2.08	0.174
120	0.2	**1.81**	**0.137**
	0.4	1.99	0.152
	0.6	2.04	0.156

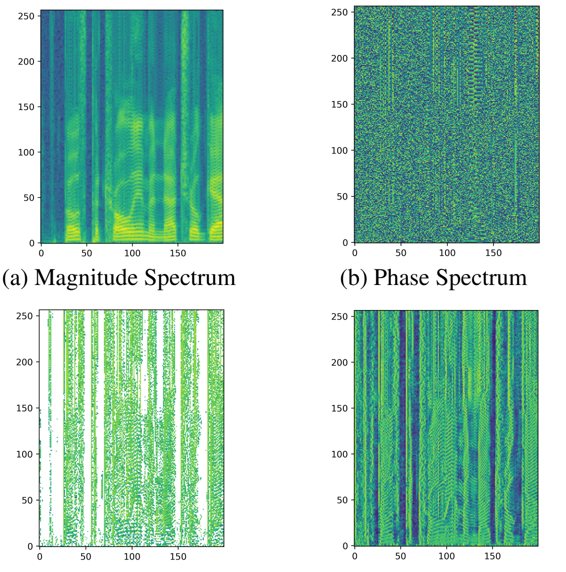
In order to examine the LearnGD feature effectiveness under different hyper-parameter settings, we report SV results in Table 1, where the LearnGD feature is computed with different window lengths and α values. We test the system performance under different hyper-parameter settings as shown in Table 1. It is noted that with the window length growing, the discriminative power of the speaker embeddings is significantly improved. This suggests that a long view of smoothing layer enables the neural network to learn a long-term representation that is potentially useful for speaker recognition, while insensitive to unwanted variability (e.g. noise, channel). Moreover, the hyper-parameter α plays a key role in controlling rang. When α varies from 0.2 to 0.6, the system performance fluctuates remarkably and gets saturated at $\alpha = 0.2$ when the length is set to 120.

4.4. Comparison with start-of-the-art systems

The comparison of our proposed LearnGD feature based SV systems to state-of-the-art systems using various input features is shown in Table 2. We observe that with the same Fast-ResNet34 feature extractor, the real and imaginary based system outperforms the magnitude spectrum based system by 0.2% EER. Compared to phase spectrum, group delay achieves a relative improvement of 16% in EER. This means that besides the widely used magnitude-based feature, the phase-based features can also encode some speaker-related information. Replacing the MODGD with the LearnGD, 34.4% relative improvement is achieved. This suggests that, by providing a flexible bias to the task, the LearnGD may have the potential to emphasize the speaker-related components and suppress the irrelevant parts in

Table 2: *Results on the Voxceleb1 dataset and extended test sets. All the methods in this table use the same training data. N/R : Not report results. PL-S: Prototypical loss and softmax loss, AP: Angular Prototypical, DAM-S: Dynamic Additive Margin Softmax loss.*

Front-end Model	Input Feature	Loss	VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H	
			EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
Thin-ResNet34[24]	Magnitude Spectrum	Softmax	2.87	N/ R	2.95	N/R	4.93	N/R
RawNet2 [25]	Raw waveform	Softmax	2.48	N/R	2.57	N/R	4.89	N/R
Fast-ResNet34[18]	Mel-FBank	AP	2.22	0.176	N/R	N/R	N/R	N/R
Fast-ResNet34[26]	Mel-FBank	PL-S	1.94	0.210	N/R	N/R	N/R	N/R
ResCNN[27]	Magnitude Spectrum	DAM-S	1.94	N/R	2.14	N/R	3.70	N/R
Fast-ResNet34	Magnitude Spectrum	AP	2.51	0.191	2.55	0.194	4.89	0.323
Fast-ResNet34	Real + Imaginary	AP	2.34	0.178	2.34	0.171	4.40	0.278
Fast-ResNet34	Phase Spectrum	AP	4.38	0.333	4.35	0.321	8.02	0.492
Fast-ResNet34	Group Delay	AP	3.68	0.232	3.51	0.255	6.48	0.389
Fast-ResNet34	MODGD	AP	2.76	0.215	2.66	0.197	5.17	0.336
Fast-ResNet34	LearnGD	AP	**1.81**	**0.137**	**1.83**	**0.132**	**3.53**	**0.228**
Fusion	LearnGD + Mel-FBank	-	1.36	0.099	1.39	0.097	2.71	0.173



(a) Magnitude Spectrum (b) Phase Spectrum

(c) Group Delay (d) Learnable Group Delay

Figure 2: *Spectrogram of different features.*

group delay domain. Thus, the feature extractor can generate more discriminative speaker representation. Moreover, the fusion of Mel-fbank and LearnGD features results in a huge improvement and achieves the start-of-the-art performance.

Furthermore, in order to verify whether the LearnGD has the ability to convey effective characteristics in spectro-temporal feature extraction, the time-frequency representations of phase, magnitude, group delay, and LearnGD are shown in Figure 2. Note that compared to phase and group delay, the LearnGD has a better resolution in the peaks and valleys region, and closely resembles the magnitude spectrum. This indicates that LearnGD may not only contain the phase information but also carry some magnitude information, thus have the power to convey meaningful attributes.

4.5. Speech recognition performance on TIMIT dataset

In this section, we evaluate the phone recognition system with proposed LearnGD feature on TIMIT dataset in terms of PER.

Table 3: *Phoneme Error Rate (PER) of different input features based phone recognition systems on TIMIT dataset.*

	Feature	**PER(%)**
Mag-based feature	Mel-FBank	18.2
Phase-based feature	Phase Spectrum	76.0
	Group Delay	22.6
	MODGD	19.9
	LearnGD	**19.4**

The results are shown in Table 3. The performance of system using original phase spectrum feature is the worst. This indicates that the ASR system cannot extract meaningful information from chaotic phase spectrum. It is noted that the GD-based system outperforms the original phase-based system by a relative 70.26% PER reduction. This indicates that the tractable GD feature is more advantageous in ASR feature extraction. By replacing the MODGD with LearnGD, a further improvement is achieved (19.9% v.s. 19.4%). This suggests that optimized for phoneme classification task, the LearnGD based system has potential to obtain more discriminative representation.

5. Conclusion

This work proposes novel phase spectrum based features, named Learnable Group Delay (LearnGD), for enhancing the performance of speaker verification. Following the definition of group delay to unwrap the phase, the proposed LearnGD features suppress the unexpected spikes with a convolutional smoothing layer. Also, to compress the value range of group delay function to amplify the low-value region, a hyper-parameter is introduced in LearnGD computation process. Both speaker verification experiments conducted on VoxCeleb2 corpus and phone recognition experiments on TIMIT demonstrate the superiority of proposed LearnGD features over widely used acoustic features, such as magnitude spectrum, Mel-Fbank *etc*.

6. 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”.

7. 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 Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

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

[4] 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.

[5] J. Peng, R. Gu, Y. Zou, and W. Wang, "Speaker-discriminative embedding learning via affinity matrix for short utterance speaker verification," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 314–319.

[6] Guangji Shi, M. M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1867–1874, 2006.

[7] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] A. Dutta, G. Ashishkumar, and C. V. R. Rao, "Phase based spectro-temporal features for building a robust asr system," *Proc. Interspeech 2020*, pp. 1668–1672, 2020.

[9] E. Loweimi, "Robust phase-based speech signal processing from source-filter separation to model-based robust asr," Ph.D. dissertation, University of Sheffield, 2018.

[10] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2011.

[11] M. J. Alam, P. Kenny, and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances," *Proc. Interspeech 2015*, 2015.

[12] K. S. R. Murty *et al.*, "Importance of analytic phase of the speech signal for detecting replay attacks in automatic speaker verification systems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6306–6310.

[13] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[14] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–125.

[15] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech communication*, vol. 49, no. 3, pp. 159–176, 2007.

[16] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," *IEEE Access*, vol. 8, pp. 161 981–162 003, 2020.

[17] R. Gu and Y. Zou, "Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation," *arXiv preprint arXiv:2001.00391*, 2020.

[18] 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.

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

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[21] J. S. Chung, J. Huh, and S. Mun, "Delving into voxceleb: Environment invariant speaker recognition," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 349–356.

[22] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6465–6469.

[23] 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.

[24] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech —& Language*, vol. 60, p. 101027, 2020.

[25] 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.

[26] S. M. Kye, J. S. Chung, and H. Kim, "Supervised attention for speaker recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 286–293.

[27] D. Zhou, L. Wang, K. A. Lee, Y. Wu, M. Liu, J. Dang, and J. Wei, "Dynamic Margin Softmax Loss for Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3800–3804.