INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge

Jee-weon Jung, Hye-jin Shim*, Hee-Soo Heo, and Ha-Jin Yu†*

School of Computer Science, University of Seoul, South Korea

jeewon.leo.jung@gmail.com, shimhz6.6@gmail.com, zhasgone@naver.com, hjyu@uos.ac.kr

Abstract

In this study, we concentrate on replacing the process of extracting hand-crafted acoustic feature with end-to-end DNN using complementary high-resolution spectrograms. As a result of advance in audio devices, typical characteristics of a replayed speech based on conventional knowledge alter or diminish in unknown replay configurations. Thus, it has become increasingly difficult to detect spoofed speech with a conventional knowledge-based approach. To detect unrevealed characteristics that reside in a replayed speech, we directly input spectrograms into an end-to-end DNN without knowledge-based intervention. Explorations dealt in this study that differentiates from existing spectrogram-based systems are twofold: complementary information and high-resolution. Spectrograms with different information are explored, and it is shown that additional information such as the phase information can be complementary. High-resolution spectrograms are employed with the assumption that the difference between a bona-fide and a replayed speech exists in the details. Additionally, to verify whether other features are complementary to spectrograms, we also examine raw waveform and an i-vector based system. Experiments conducted on the ASVspoof 2019 physical access challenge show promising results, where t-DCF and equal error rates are 0.0570 and 2.45 % for the evaluation set, respectively.
Index Terms: replay detection, anti-spoofing, speaker recognition, representation learning, deep neural networks

1. Introduction

Automatic speaker verification (ASV) systems are being widely applied to various industries. However, spoofing attacks are becoming a threat to the reliability of ASV systems, necessitating the study of spoofing detection systems. Following this trend, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) initiative is providing a platform for researches to follow-up, study, and compare spoofing detection systems. The ASVspoof Challenge has covered various kinds of spoofing attacks, such as text-to-speech (TTS) and voice conversion (VC) in 2015, and replay attacks in 2017 [1, 2]. ASVspoof2019 challenge deals with advances in TTS and VC technology as logical access and controlled simulation of replay attack as physical access [3] VC and TTS require expertise and specialized equipment. In contrast, replay attacks does not require any expertise nor specialized equipment. It can be simply conducted by acquiring target speaker's voice using a recording device, and then replaying using a playback device. In this process, a different combination of replay and playback device with background environment can be used which is referred to as 'replay configuration'. Despite the simplicity of attack scheme, replay attack has been proved as an effective way to deceive an ASV system. This study concentrates on replay detection task.

Through a survey on previous studies in replay detection including past ASVspoof competitions, we found that a number of researches have focused on finding discriminative features to improve spoofing detection [4–7]. Such features include constant Q cepstral coefficients (CQCC), inverse Mel-filter cepstral coefficients (IMFCC), linear prediction cepstral coefficients (LPCC), and group delay (GD)-grams. These features concentrate on representing the characteristics of a speech, that is considered discriminative in conventional knowledge for replay detection. For instance, IMFCC concentrates on high frequency bands, utilizing the knowledge that high frequency bands in replayed speech are often distorted. However, as a result of advances in both recording and playback devices, distortion that reside in a replayed speech diminishes. We hypothesize that because of this phenomenon, discriminative power of conventional features will decrease.

To deal with decreasing distortions in replayed speech, we explore an approach of minimizing the intervention of conventional knowledge and fully exploit DNN-based data driven approach. Our main focus in this study is to provide appropriate unprocessed, complementary information with high-resolution to facilitate end-to-end DNN. Complementary information combining not only general spectrograms which include magnitude information, but also phase information and power spectral density (PSD) is explored. We explore phase information, which has been shown to be effective in replay attack detection [6, 8–10], with PSD for concentrating on the distribution of the power signal over frequencies rather than concentrating on spectral contents. To verify the effectiveness, we investigate model-level and score-level ensembles of various spectrograms with PSD. Experiments confirm that using complementary features contributes in the direct modeling of spectrogram-based deep neural networks (DNNs).

Furthermore, we used high-resolution of 2048 fast Fourier transform (FFT) bins for all features. The purpose is being able to represent the subtle difference between bona-fide speech and spoofed speech. Because of the advancement of replay attacks, the difference may be more subtle, and less obvious, requiring focus on minute distinctions. Our comparative experiments show that the resolution significantly affects actual performance (see Table 3).

2. End-to-end DNN

We introduce an end-to-end DNN that is used to deal with decreasing distortions in replayed speech by minimizing the intervention of conventional knowledge. Using spectrograms as input, end-to-end DNN replaces the sub-process of selecting dis-

*These authors contributed equally.

† Corresponding author

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2017R1A2B4011609)

http://dx.doi.org/10.21437/Interspeech.2019-1991

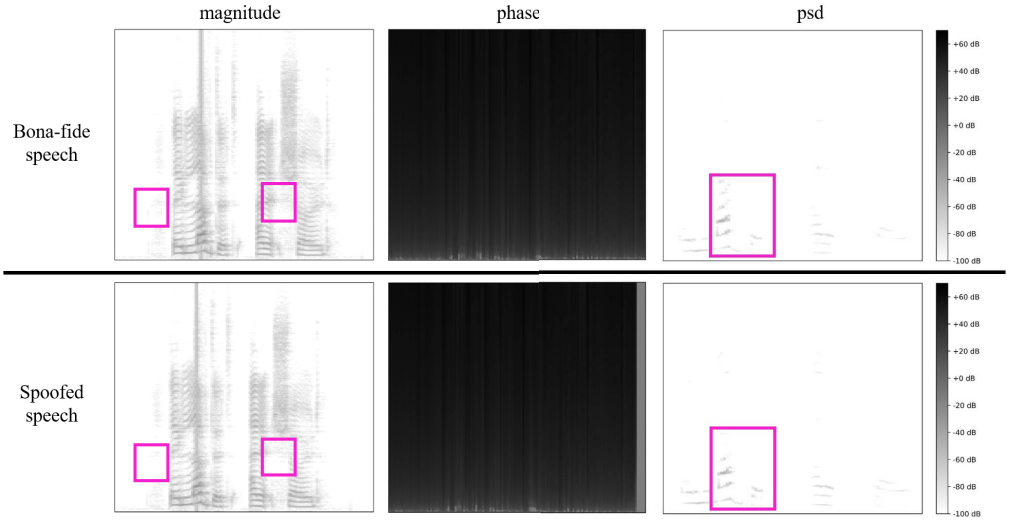

Figure 1: *Visualization of bona-fide (upper) and replayed (lower) spectrograms and PSD: magnitude (left), phase (mid), and PSD (right). Minor differences in small regions (pink boxes) demonstrate the difficulty of replay attack spoofing detection task and the necessity of high-resolution.*

criminative parts, which makes the intermediate representations adaptable to the data. An output of this model directly indicates a decision score when spectrograms are input, which simplifies the process pipeline. The state-of-the-art in various audio domain tasks has adopted an approach that utilizes a DNN that directly inputs spectrograms [11–14].

The DNN used in this study comprises convolutional neural networks (CNNs), gated recurrent units (GRUs) and fully connected layers (CNN-GRU) as used in [15–17]. In this architecture, input features are first processed using convolutional layers to extract frame-level embeddings. Convolutional layers comprise residual blocks [18] with identity mapping [19] to facilitate the training of deep architectures. Specifically, the first convolutional layer of our model processes local adjacent time and frequency domains and is gradually aggregated by repeating pooling operations to extract frame-level embedding. Then, a GRU layer is employed to aggregate extracted frame-level features into a single utterance-level feature. One fully connected layer is used to transform the utterance-level feature. An output layer with two nodes indicates either the input utterance is bona-fide or spoofed.

3. Complementary high-resolution feature

In this section, we introduce the key aspects to facilitate training end-to-end DNN without intervention based on human priors; providing complementary information and using high-resolution. To generalize towards unknown replay configurations, we hypothesize that the approach of providing varied, raw information as input and performing data-driven feature selection with DNN would provide a more appropriate process of feature extraction for spoofing detection. Based on this hypothesis, spectrograms including various information with high-resolution are explored in expectation of outperforming acoustic features extracted with conventional knowledge. Specifically, phase information and PSD was exploited in addition to general spectrograms including magnitude information. In general, a spectrogram refers to a magnitude spectrogram containing absolute values of the fast Fourier transform (FFT). Whereas phase information was often overlooked in many audio domains, recent studies have demonstrated that phase-based features provide discriminative information for replay detection [6, 7, 20]. We use magnitude spectrogram and phase spectrogram supposing that phase information would supplement the magnitude information without requiring an additional extraction process, as both use partial information of FFT. We also exploit PSD for further improvement. Because PSD concentrates on the distribution of signal power over frequencies, it differs from magnitude or phase, which concentrate on spectral contents. Concurrently using PSD and spectrograms therefore enables consideration of the frequency distribution of the overall signal strength as well as the harmonics of amplitude and phase in the signal. To exploit this diverse information, we explore the combinations in both the model-level and score-level. Model-level ensemble inputs various features to a single DNN, whereas score-level ensemble exploits multiple DNNs and conducts score summation of DNNs' outputs, respectively. To further analyze the relations between complementary information, we compared each combination using different spectrograms. Figure 1 shows that, even with the same speech, different clues (pink boxes) for spoofing detection can be presented, depending on the types of spectrograms used.

As noted previously, with the advance in quality of audio devices, the difficulty of detecting spoofed utterances has been increased, as prominent differences between bona-fide speech and spoofed speech have been reduced. This necessitates the usage of high-resolution input that can be used to demonstrate the subtle differences residing in replay spoofed utterances. A high-resolution of 2,048 FFT bins was used for all spectrograms in this study. We were inspired by the experiments in *Tom et al.* [6] that attention-based GD-grams significantly outperformed spectrograms. The GD-grams used in the *Tom et al.* [6] are obtained using 2,048 FFT bins, which had higher resolution than spectrograms. We hypothesized that the difference in resolution could have also conducted a key role to the performance besides the difference of used feature. To verify this hypothe-

Table 1: *DNN architecture (l: length of input sequence).*

layer	output shape	kernel size	stride
Conv1	$l \times 1024 \times 16$	3×7	1×1
Res1	$(l/2) \times 257 \times 32$	3×5	2×4
Res2	$(l/4) \times 65 \times 64$	3×5	2×4
Res3	$(l/8) \times 17 \times 128$	3×5	2×4
Pool	$(l/8) \times 1 \times 128$	1×17	1×17
GRU	1×512	-	-
Dense1	64	512×64	-
Output	2	64×2	-

sis, we conducted an comparative experiment. Results shown in Table 3 match our hypothesis where equal error rate (EER) of a spectrogram using 2,048 FFT bins significantly outperformed an identical system with 512 FFT bins.

4. Experimental settings

DNN training was implemented using Keras, a deep learning library for python, with a Tensorflow backend [21–23]. The i-vector extraction was conducted using the Kaldi toolkit [24].

4.1. Dataset

We used the ASVspoof 2019 physical access dataset for all experiments. This dataset comprises 54,000 utterances as the training set, 29,700 utterances as the development set, and 137,457 utterances as the evaluation set. Utterances are recorded from 20 speakers (8 male, 12 female) at a 16-kHz sampling rate with 16-bit resolution. Training and development data comprises 27 different acoustic configurations using 3 room sizes, 3 levels of reverberation, and 3 speaker-to-ASV microphone distances. 9 different replay configurations are used, as combinations of 3 categories of attacker-to-talker recording distances and 3 categories of loudspeaker quality. The acoustic and replay configurations of the evaluation set are different from those of the training and development set.

4.2. Spectrograms, raw waveforms, and i-vector extraction

Spectrograms were extracted using a hamming window with a length of 50 ms and a shift size of 20 ms. Representation using magnitude, phase spectrogram, and PSD were extracted with 2,048 FFT bins each. The number of the time axis was fixed to 120 ($\approx$ 2.4 s), by either cropping long utterances or duplicating short utterances at the training phase for batch construction, depending on their lengths. Whole utterances were input at the evaluation phase without duration adjustment.

Raw waveforms were directly input to the DNN without any pre-processing. The pre-emphasis layer was excluded, which differs from the setup in [16], based on a comparison experiment. For batch construction, the length of each utterance was fixed to 26,244 samples ($\approx$ 1.64 s) by either performing random cropping for long utterances or duplicating for short utterances. At the evaluation phase, whole utterances were input to the DNN.

The i-vectors were extracted using a universal background model with 256 diagonal Gaussian components which input 20-dimensional Mel-frequency cepstral coefficients with its first and second derivatives, comprising 60-dimensional acoustic features. 200-dimensional i-vectors were extracted, and neither linear discriminant analysis nor length normalization were applied.

Table 2: *Performance comparison between spectrogram-based systems with different types, raw waveform, and i-vector on the development set. Spectrogram-based with more than one type shows model-level ensemble results.*

System	**t-DCF**	**EER (%)**
Baseline (CQCC-GMM)	0.1953	9.87
Spec-magnitude	**0.0482**	**1.76**
Spec-psd	0.1153	3.74
Spec-phase	0.2145	8.04
Raw waveform	0.1915	8.03
i-vector	0.2119	8.74
Spec-magnitude&psd	**0.0491**	**1.75**
Spec-magnitude&phase	0.0590	2.11
Spec-psd&phase	0.1159	3.91
Spec-magnitude&psd&phase	0.0688	2.11
Spec-score-level ensemble	**0.0306**	**1.05**

4.3. DNN architecture

A slightly modified ResNet was used for modeling the spectrograms, accounting for different stride sizes for time and frequency domains due to high-resolution in the frequency domain, and the number of residual blocks was adjusted to fit the provided ASV2019 physical access dataset. The raw waveform CNN-GRU model, proposed in [17], was used with a few modifications: one less residual block, a different specified input utterance length at training phase to fit the dataset, and additional loss functions for training (center loss [25] and speaker basis loss [26]). This model first extracts 128-dimensional frame-level representations using 1-dimensional convolutional layers. Then a GRU layer with 512 nodes combines the extracted frame-level features into utterance-level features.

A simple fully-connected DNN with 3 layers, each with 1,024 nodes, was used for i-vector modeling. For all DNNs, he normal initialization [27], weight decay with $\lambda = 1e^{-4}$ was applied and trained with AMSGrad optimizer [28]. Additionally, for all systems, the output layer has two nodes, each indicating bona-fide and spoofed utterances. The output layer's node value that indicates a bona-fide utterance was directly used as the score (in end-to-end fashion) without additional modeling when an utterance was input. The DNN architecture is summarized in Table 1[12].

5. Result analysis

In this section, we first evaluate the single systems, then verify the effect of using complementary features in model-level and score-level, and then demonstrate that high-resolution is necessary. First, the evaluations of single systems are shown in the 2^{nd} to 6^{th} rows of Table 2. All single systems clearly outperform the CQCC baseline. Magnitude spectrogram, which uses the absolute value of FFT, seems most appropriate for replay attack spoofing detection.

Second, the effect of using complementary spectrograms is analyzed. The results of ensemble systems in model-level and score-level fusion are shown in the 7^{th} to 10^{th} rows and the

[1]Implementation of raw waveform processing, spectrograms extraction, and DNN architecture are in `https://github.com/\Jungjee/ASV2019_competition_Jung`

[2]Modified model in PyTorch with training script is in `https://github.com/Jungjee/ASVspoof2019_PA`

Table 5: *Performance comparison on the evaluation set using various attacker-to-talker distances and loudspeaker quality of the CQCC baseline and our submitted primary system. Two sets of labels refer to attacker-to-talker distance (A: 10-50 cm, B: 50-100 cm, C: far than 100 cm) and loudspeaker of quality (A: perfect, B: high, C: low) respectively.*

Metric	System	**Pooled**	AA	AB	AC	BA	BB	BC	CA	CB	CC
t-DCF	CQCC-baseline	0.2454	0.4975	0.1751	0.0529	0.4658	0.1483	0.0433	0.5025	0.1360	0.0461
	Primary	0.0570	0.1603	0.0416	0.0207	0.0839	0.0232	0.0111	0.0529	0.0184	0.0081
EER	CQCC-baseline	11.04	25.28	6.16	2.13	21.87	5.26	1.61	21.10	4.70	1.79
	Primary	2.45	6.65	1.68	0.82	3.33	0.90	0.45	2.17	0.63	0.30

Table 3: *Performance comparison of various FFT resolutions. Magnitude spectrogram, single best system, was used for comparison. In these experiments, window length and shift size were fixed to 30 ms and 10 ms respectively to ensure that the number of samples within a window is greater than nFFT. The performance difference of the nFFT 2,048 model with that of Table 3 is due to the different window length and shift size.*

System	**nFFT**	**t-DCF**	**EER (%)**
Spec-magnitude	**2048**	**0.0894**	**3.07**
	1024	0.1226	3.81
	512	0.2488	7.83

11^{th} row of Table 2, respectively. Model-level did not show improvement, but score-level resulted in significant improvement. Surprisingly, including model-level ensemble systems in score-level ensemble additionally brought further performance improvement where score-level ensemble of 7 spectrogram-based systems demonstrated an EER of 1.05 %. To verify if other features can also complement various high-resolution spectrograms, we explored two more features. We explored raw waveform and i-vector because raw waveform does not include any pre-processing, and i-vector is a well-known utterance-level representation extracted based on human knowledge.

Next, Table 3 demonstrates the necessity of high-resolution by comparing performance with different numbers of FFT bins. Results show that high-resolution features are indeed critical for replay attack detection. Additionally, by comparing magnitude spectrogram systems with 2,048 FFT bins in Table 2 and Table 3, there was a considerable performance difference between spectrograms with a 50-ms window and a 20-ms shift in comparison to a 30-ms window and a 10 ms shift, where EERs were 1.76 % and 3.07 % respectively. Through this result, we note that the window length and shift size are also crucial for replay attack detection, as reported in [29].

Table 4 shows the results of submitted systems for the ASV2019 physical access challenge. The magnitude spectrogram based system was submitted as Single. Based on the improvement brought with score-level ensemble, the primary system comprises spectrogram, raw waveform, and i-vector based models. The combined i-vector and 7 spectrograms based model was submitted as Contrastive1, and the combined raw waveform and 7 spectrograms based model was submitted as Contrastive2. Adding both raw waveform and i-vector further reduced EER to 0.96 %, and was submitted as the Primary system for the competition.

Performance analysis of the baseline CQCC system and the 'Primary' submission on different replay configurations, mainly for attacker-to-talker distance and replay device quality, is addressed in Table 5. Results demonstrate that the proposed system clearly outperforms the baseline regardless of replay configurations in terms of both t-DCF and EER. Although both attacker-to-talker distance and replay device quality affected the performance significantly, our 'Primary' system was more robust towards replay spoofing using high quality devices. For replay attacks using high quality device (compare AA, BA, and CA), the baseline system consistently exhibited EER higher than 20 % where our 'Primary' submission could show improved performance as attacker-to-talker distance decreased. We interpret that using high-resolution played a key role for this result.

Table 4: *t-DCF and EERs for the submissions to the ASV2019 physical access challenge condition development and evaluation set.*

Submission	**t-DCF**		**EER (%)**	
	val	**eval**	**val**	**eval**
Primary (7 spec+wave+i-vec)	**0.0244**	**0.0570**	**0.96**	**2.45**
Single (spec-mag)	0.0482	0.1255	1.76	4.79
Contrastive1 (7 spec+i-vec)	0.0246	0.0692	0.98	2.81
Contrastive2 (7spec+wave)	0.0284	0.0632	1.10	2.73

6. Conclusion

In this study, we focus on replacing the hand-crafted feature extraction process by directly modeling spectrograms using DNNs in end-to-end fashion. As advanced recording and playback devices arise, characteristics of a speech, considered discriminative in conventional knowledge for replay detection diminish. Thus, it has become increasingly difficult to distinguish bona-fide speech from spoofed speech. To detect unrevealed characteristics that reside in a replayed speech, we directly input spectrograms into an end-to-end DNN without knowledge-based intervention. Utilizing explorations of this study such as complementary information and high-resolution further facilitates a data-driven approach. Additionally, the ensemble use of different features, including raw waveform and i-vector, was verified to further increase performance. The primary system submitted to the ASV2019 challenge demonstrated a t-DCF of 0.0570 and an EER of 2.45 % and was compared to a t-DCF of 0.2454 and an EER of 11.04 % baseline CQCC-GMM on the ASV2019 physical access challenge evaluation set.

7. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[4] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge." in *Interspeech*, 2017, pp. 7–11.

[5] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Independent modelling of high and low energy speech frames for spoofing detection." in *INTERSPEECH*, 2017, pp. 2606–2610.

[6] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," *INTERSPEECH, Hyderabad, India*, 2018.

[7] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for asvspoof 2015 challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] K. Srinivas and H. A. Patil, "Relative phase shift features for replay spoof detection system," 2018.

[9] D. Li, L. Wang, J. Dang, M. Liu, Z. Oo, S. Nakagawa, H. Guan, and X. Li, "Multiple phase information combination for replay attacks detection," *Proc. Interspeech 2018*, pp. 656–660, 2018.

[10] T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, and J. Epps, "Detection of replay-spoofing attacks using frequency modulation features," *Proc. Interspeech 2018*, pp. 636–640, 2018.

[11] L. Wan, Q. Wang, A. Papir, and I. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[12] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[13] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[14] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[15] J. Jung, H. Heo, I. Yang, H. Shim, and H. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5349–5353.

[16] ——, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," in *Proc. Interspeech 2018*, 2018, pp. 3583–3587.

[17] J.-w. Jung, H.-s. Heo, H.-j. Shim, and H.-j. Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," *arXiv preprint arXiv:1810.10884*, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 630–645.

[20] F. Auger, É. Chassande-Mottin, and P. Flandrin, "On phase-magnitude relationships in the short-time fourier transform," *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 267–270, 2012.

[21] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[22] A. Martín, A. Ashish, B. Paul, B. Eugene *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015. [Online]. Available: http://download.tensorflow.org/paper/whitepaper2015.pdf

[23] A. Martin, B. Paul, C. Jianmin, C. Zhifeng, D. Andy, D. Jeffrey, D. Matthieu, G. Sanjay, I. Geoffrey, I. Michael, K. Manjunath, L. Josh, M. Rajat, M. Sherry, M. G. Derek, S. Benoit, T. Paul, V. Vijay, W. Pete, W. Martin, Y. Yuan, and Z. Xiaoqiang, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[26] H.-S. Heo, J.-w. Jung, I.-H. Yang, S.-H. Yoon, H.-j. Shim, and H.-J. Yu, "End-to-end losses based on speaker basis vectors and all-speaker hard negative mining for speaker verification," *arXiv preprint arXiv:1902.02455*, 2019.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[28] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," 2018.

[29] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, S. Marcel, H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 11, pp. 2098–2111, 2017.