Fake Speech Detection Using Residual Network with Transformer Encoder

Zhenyu Zhang
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China
zhangzhenyu@iie.ac.cn

Xiaowei Yi
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China
yixiaowei@iie.ac.cn

Xianfeng Zhao*
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China
zhaoxianfeng@iie.ac.cn

ABSTRACT

Fake speech detection aims to distinguish fake speech from natural speech. This paper presents an effective fake speech detection scheme based on residual network with transformer encoder (TE-ResNet) for improving the performance of fake speech detection. Firstly, considering inter-frame correlation of the speech signal, we utilize transformer encoder to extract contextual representations of the acoustic features. Then, a residual network is used to process deep features and calculate score that the speech is fake. Besides, to increase the quantity of training data, we apply five speech data augmentation techniques on the training dataset. Finally, we fuse the different fake speech detection models on score-level by logistic regression for compensating the shortcomings of each single model. The proposed scheme is evaluated on two public speech datasets. Our experiments demonstrate that the proposed TE-ResNet outperforms the existing state-of-the-art methods both on development and evaluation datasets. In addition, the proposed fused model achieves improved performance for detection of unseen fake speech technology, which can obtain equal error rates (EERs) of 3.99% and 5.89% on evaluation set of FoR-normal dataset and ASVspoof 2019 LA dataset respectively.

CCS CONCEPTS

• **Security and privacy** → *Authentication*; • **Computing methodologies** → *Artificial intelligence*; *Natural language processing*.

KEYWORDS

Fake speech detection, transformer encoder, residual network, data augmentation, logistic regression, score-level fusion

ACM Reference Format:
Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. 2021. Fake Speech Detection Using Residual Network with Transformer Encoder. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '21), June 22–25, 2021, Virtual Event, Belgium.*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3437880.3460408

1 INTRODUCTION

Over the recent years, the growth of online social media and speech signal processing techniques has greatly facilitated emerging of a lot of fake speeches [21]. In addition, the state-of-the-art voice conversion (VC) and text-to-speech (TTS) systems achieve such a high level of naturalness that even humans have difficulties to distinguish fake speech from natural speech. These fake speeches have the potential to cause serious problems in society, for example, the security of automatic speaker verification (ASV) systems is compromised by fake speech attacks [4]. Meanwhile, much of the works within fake speech detection have been studied for protecting ASV systems from fake speech attacks [8].

The popular approaches for fake speech detection consist of front-end features and back-end classifiers. The target of front-end features is capturing the artifacts introduced during tampering speech signal. Aleksandr *et al*. [22] proposed i-vector based representation of speech clips, which was promising to provide generalized fake speech detection. Sahidullah *et al*. [20] made a comparative study and demonstrated the efficacy of various short-term power spectrum and phase features for synthetic speech detection. Monisankha *et al*. [16] combined linear frequency cepstral coefficients (LFCCs) and inverted Mel-frequency cepstral coefficients (IMFCCs) as an excellent discriminative feature for fake speech detection. Massimiliano *et al*. [25] proposed short-term spectral features like constant Q cepstral coefficients (CQCCs) and exhibited best results on ASVspoof 2015 dataset. The winning system of ASVspoof 2017 challenge adopted a combination of standard Mel-frequency cepstral coefficients (MFCCs) and cochlear filter cepstral coefficients (CFCCs) to detect fake speeches [17]. Chen *et al*. [2] proposed a robust representation based on deep neural network to extract the representative acoustic features for the synthetic speech detection.

Apart from focusing on front-end acoustic features, some researchers pay keen attention on developing effective back-end classifiers. Wu *et al*. [31] introduced Gaussian mixture models

[*] Corresponding author

IH&MMSec '21, June 22–25, 2021, Virtual Event, Belgium.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8295-3/21/06…$15.00
https://doi.org/10.1145/3437880.3460408

(GMMs) as back-end classifiers to detect fake speeches. Sonawane *et al.* [23] used support vector machine (SVM) to classify natural and fake speech. Although the traditional machine learning models have achieved good results, the deep neural networks are also applied as back-end classifiers. Galina *et al.* [11] developed light convolutional neural network (LCNN) architecture by using max-feature-map activation, and then investigated the efficiency of angular margin based softmax activation function for training robust deep LCNN classifier to solve the synthetic speech detection tasks in [12]. Jung *et al.* [7] directly inputted spectrograms into an end-to-end deep neural network (DNN) which comprised convolution neural network and gated recurrent units (CNN-GRUs). Alzantot *et al.* [7] built three variants of the residual convolutional neural network (ResNet) with different acoustic feature representations respectively for synthetic speech detection. Bhusan *et al.* [3] used five deep neural network models and two shallow models as single classifier model to calculate the score of each speech clip, and then the ensemble model was used to combine information from different classifier models, which achieved the third position of the ASVspoof 2019 challenge.

In this paper, we present an effective fake speech detection scheme based on data augmentation techniques and residual network with transformer encoder (TE-ResNet). In the proposed scheme, we first obtain five times the amount of training dataset via various speech data augmentation techniques. Then, the spectrum features are extracted from the speech clips and sent to transformer encoder to extract the deep features. Next, the residual network is used to further process the deep features and calculate the score that the speech is fake. Finally, the score from several different detectors are fused to compensate the shortcomings of each single model. The experiment results demonstrate that our proposed single detection model outperforms the existing fake speech detectors, and the performance is also improved by score-level fusion strategy. To allow researchers to verify, reproduce, and extend our work, we provide the detection systems with corresponding detection results as an open source on GitHub (https://github.com/Amforever/TE-ResNetandDAforFSD).

The remaining part of the paper is organized as follows. The preliminaries are introduced in Section 2 and the architecture of our scheme is presented in Section 3. Experimental datasets and evaluation measures are described is Section 4. The overall performances of fake speech detection methods on two datasets are described in Section 5. Section 6 draws conclusions and directions for future work.

2 PRELIMINARIES

2.1 Acoustic Features

The front-end acoustic features are pretty important for efficiently detecting fake speech. In the proposed scheme, we adopt log power spectrum (LPS), Mel-frequency cepstrum coefficient (MFCC) and constant Q cepstral coefficient (CQCC) as the front-end acoustic features respectively. These spectrum features are better input representations than time-domain waveforms for efficiently training the back-end classifiers. In addition, through various time-frequency representations such as spectrograms, we can get a rich representation of the temporal and spectral structure of the speech signal. The short-time Fourier transform (STFT) and the constant Q transform (CQT) are popular tools to specify the time-domain signal into the time-frequency representations.

The STFT performs a Fourier transform on a short segment which is extracted from a longer speech signal upon its multiplication with a suitable window function. The sliding window is applied repetitively in order to analyse the local frequency content of the longer speech signal as a function of time [15]. LPS and MFCC are derived from STFT and act as the front-end acoustic features for the fake speech detection approaches. The LPS of a speech signal $x(t)$ is calculated as:

$$LPS(w)=log|F\left(x\left(t\right)\right)|^2, \tag{1}$$

where F represents the Fourier transform and $|\cdot|^2$ operation computes a component-wise squared magnitude.

The MFCC of a speech signal $x(t)$ is typically extracted according to:

$$MFCC(q)=\sum_{m=1}^{M} log|MF\left(m\right)|cos\left[\frac{q\left(m-1/2\right)\pi}{M}\right], \tag{2}$$

where $MF(m)$ is Mel-frequency spectrum of a discrete time domain signal $x(t)$, and M is the number of Mel-filters. $MFCC(q)$ is applied to extract a number of coefficients less than the number of Mel-filters M. Typically, M=25 and q varies between 13 and 20. The extraction procedure of MFCC is described as follows. Firstly, we compute the STFT of the speech signal using a periodic Hamming window of length 1024 and an overlap of length 256 samples. Then, we map the powers of spectrum onto the Mel-scale by Mel-filter bank. Next, discrete cosine transform (DCT) is applied to reduce the dimensionality, and the first 24 coefficients are picked as static MFCC. Finally, the delta and acceleration of static MFCC are calculated and appended to the static coefficients to form a 72-dim feature vector. When a speech clip consists of 32,000 samples, we get the acoustic feature matrix with 126 rows and 72 columns.

The CQT time-frequency analysis was proposed by Kashima *et al.* [9], where octaves are geometrically distributed while the centre frequencies of each filter are linearly spaced. Unlike the fixed time-frequency resolution of Fourier transform method, CQT can give a higher frequency resolution for lower frequencies along with a higher temporal resolution for higher frequencies. The CQCC is derived from CQT and extracted as described in [26]. The CQCC of a speech signal $x(t)$ is calculated as:

$$CQCC(p) = \sum_{l=1}^{L} log|X^{CQ}\left(l\right)|^2 cos\left[\frac{p\left(l-1/2\right)\pi}{L}\right], \tag{3}$$

where X^{CQ} is CQT of a discrete time domain signal $x(t)$. $p = 0, 1, ..., L-1$ and l is the newly resampled frequency bin. To compute CQCC matrix, after applying CQT, we calculate a power spectrum and take a logarithm. Then a uniform re-sampling is performed, followed by DCT. Finally, we get the 126 × 72 CQCC matrix which is the same shape as MFCC matrix.

2.2 Transformer Encoder

Transformer architecture is used to model of long-term dependencies in sequential data, which has shown promising results in many sequence to sequence transformation tasks recently [28].

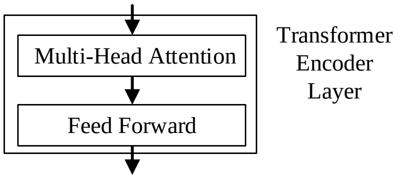

Figure 1: The structure of a encoder layer in the transformer encoder.

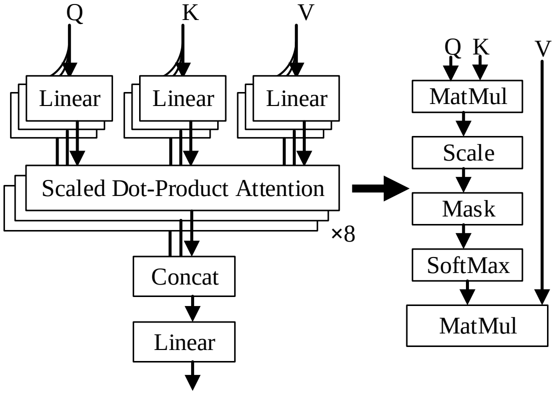

Figure 2: The detail architecture of multi-head attention, where Q, K and V represent query, key, and value respectively.

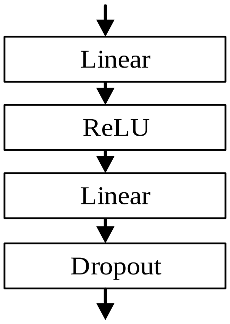

Figure 3: The detail architecture of feed-forward.

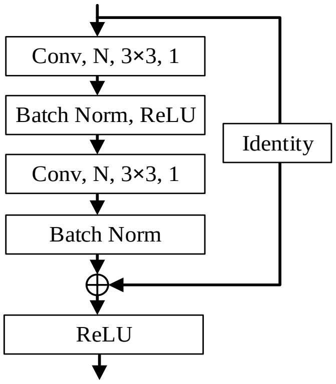

Figure 4: The detail architecture of ResNet Bock.

It improves on recurrent neural networks' most shortcoming of slow training by using the self-attention mechanism. The attention mechanism is adopted to form global dependencies between input and output, which leads the transformer model as an completely encoder-decoder architecture. The transformer encoder consists of a set of encoding layers that process the input iteratively one layer after another, which generates their representations of the input statements. The encoding layers of transformer encoder are all identical in structure, and each one is broken down into two sub-layers: multi-head attention and feed-forward. The architecture of a encoder layer in the transformer encoder is shown in Figure 1.

As in Figure 1, multi-head attention is the brain of transformer encoder model, which allows the model to jointly attend to information from different representation sub-spaces at different positions. We depict the architecture of multi-head attention block in Figure 2. The attention module repeats its computations multiple times in parallel, and each of these is called an attention head (8 heads in our model). The attention module splits its query, key, and value parameters 8 ways and passes each split independently through a separate head. All of these similar attention calculations are then combined together to produce a final attention score. The outputs of multi-head attention layer are fed to a feed-forward neural network. As described in Figure 3, the feed-forward neural network consists of two linear layers with a ReLU activation in between. The exact same feed-forward neural network is independently applied to each position following multi-head attention block.

As it is mentioned above, the transformer encoder receives a list of vectors as input. It processes the input list by passing these vectors into a multi-head attention layer, then into a feed-forward neural network. Transformer encoder can be expanded to very deep depths to fully exploit features of deep neural network models and improve accuracy. Considering inter-frame correlation of the speech signal, we aim to use transformer encoder to per-process the front-end acoustic features.

3 THE PROPOSED SCHEME

In this section, the architecture of our proposed scheme is elaborated, and each part of the neural network is analyzed in detail. The whole structure of our scheme is shown in Figure 5. As demonstrated in Figure 5, the quantity of training data is increased by five times through the speech data augmentation strategies. Then, acoustic feature matrix of speech clip with the size of 126×72 is extracted as the input data of the transformer encoder. Next, residual network with transformer encoder is used to extract higher level contextual representations of the acoustic characteristics and calculate the score that the speech is fake.

3.1 TE-ResNet

The TE-ResNet consists of two parts: transformer encoder and residual network. Transformer encoder (described in Section 2.2) is used to pre-process the acoustic features matrix to get the deep feature maps. Firstly, we use linear layer with layer norm to replace

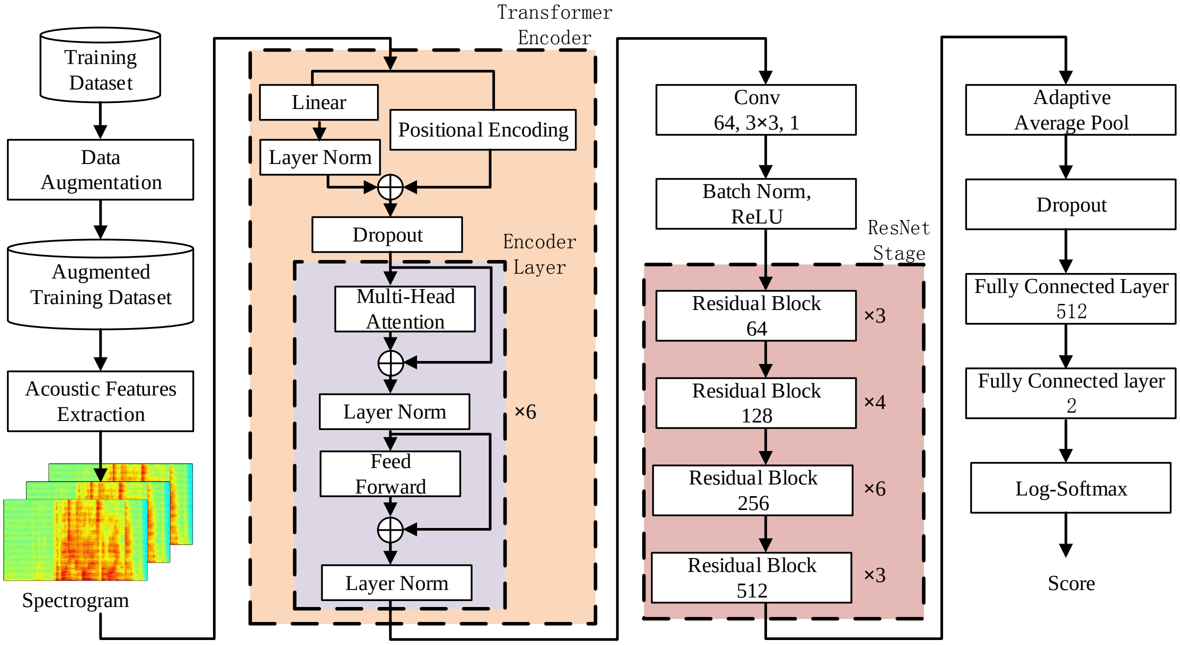

Figure 5: Structure of the proposed scheme. The parameters in Residual Block and Fully Connected Layer are the number of channels and neurons respectively. Outside the rectangular box is the number of stacked blocks. In the Conv box are the number of channels, kernel size and the stride of slide window.

input feature matrix and add it with the feature matrix position encoding, which produce position embedding matrix of shape 64 × 126 × 256 (number of batch size, number of frames per speech clip, number of expected features in the encoder). Then, we apply dropout of 0.1, and fed position embedding matrix to the query, key, and value of the first encoder layer in the stack. It is passed through a multi-head attention block (as described in Figure 2) and is added to the original input using a residual connection, and a consecutive layer normalization is applied on the sum. Finally, a small feed-forward neural network (as described in Figure 3) is followed, and is added to the original input using a residual connection, and a consecutive layer normalization is applied on the sum. We stack six of encoder layers on top of each other.

We modify architecture of the 34-layer ResNet and input the deep features obtained from transformer encoder to residual network. The residual network is used to further process the deep features and calculate the score that the speech is fake. Firstly, the convolutional layer (3 × 3 kernel, 1 stride and 64 channels), batch normalization, and ReLU activation are processed the feature map from transformer encoder. Then, a ResNet stage is followed, which consists of four kinds of Residual Blocks and each kind of Residual Block is repeated three to six times (3, 4, 6, 3) (as described in Figure 4). Finally, an adaptive average pool, a dropout layer, two fully connected layers, and a Log-Softmax layer are used for calculating the score that the speech is fake.

3.2 Speech Data Augmentation

Speech data augmentation, which is a common strategy that adopted to increase the quantity of training data, has been shown to be effective in training neural networks with better robustness. The speech data augmentation strategy has been proposed as a method to generate additional training data for automatic speech recognition (ASR) [13]. More generally, the traditional speech data augmentation techniques have achieved state-of-the-art performance in the many domains [30]. Inspired by the recent success of augmentation in the speech and vision domains, we apply five traditional speech data augmentation methods on the raw waveform. These methods are simple and computationally cheap to apply, as it directly acts on the raw speech and will produce many additional training data. Specifically, the five traditional speech data augmentation methods are Gaussian noise addition (GNA), Signal-to-noise ratio noise addition (SnrNA), time shifting, pitch shifting, and time stretching.

To show the impact of each augmentation method on speech, we drawn waveforms and spectrograms of speech clips in Figure 6 and Figure 7. The spectrogram is a visual representation of the spectrum of frequencies of a speech signal as it varies with time, which is usually depicted as a heat map with the intensity shown by varying the color or brightness. The top in the Figure 6 is the waveform and spectrogram of the original speech clip, and the middle and bottom are GNA and SnrNA added signal versions respectively. There are time shifting, time stretching and pitch shifting speech signal versions in the Figure 7. As we can see from the Figure 6 and Figure 7, the waveforms of original and augmented speech signal are almost the same expect for pitch shifting version and time stretching version. However, the spectrograms of different versions

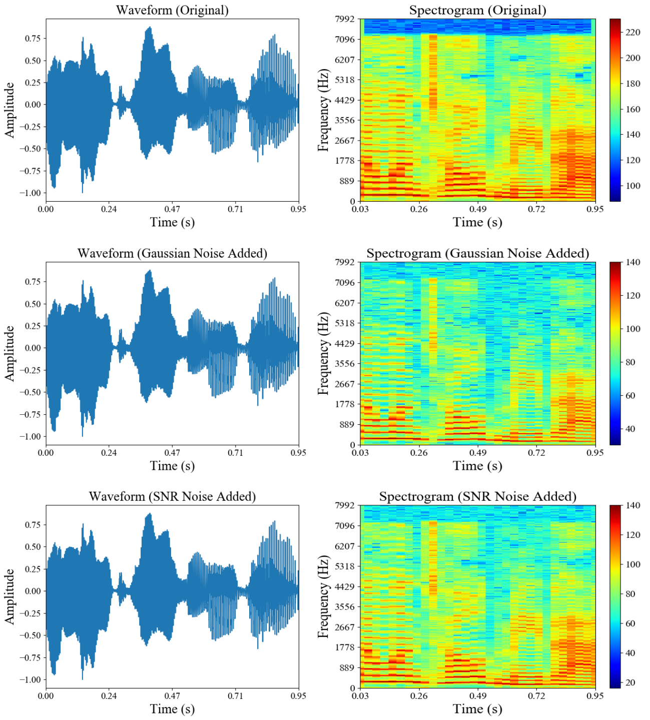

Figure 6: The waveform and spectrogram of the original speech clips and corresponding augmented speech clips.

are obviously different, which will make more variety of acoustic features to train neural networks and enhance the generalization of the neural networks. After making speech data augmentation, we have five times the amount of original training dataset.

3.3 Multi-Model Fusion Strategy

Multi-model fusion strategy integrates multiple models to boost the performance of single model, which has been used in many areas such as image processing [5] and fake speech [24]. In the fake speech detection schemes, each front-end acoustic feature LFCC, CQCC and LPS can represent specific acoustic characteristic of speech signal, and each back-end classifier can learn the representations in its own rules. To take full advantage of the acoustic features and classification abilities of neural networks, we fuse detectors' scores using the logistic regression as the final score (as described in Figure 8).

Unlike a weighted average ensemble where multiple sub-models contribute equally to a combined score, logistic regression can use the set of scores as a context and conditionally decide to weigh the input scores differently. This allows well-performing models to contribute more and less-well-performing models to contribute less. In addition, simplicity and interoperability of logistic regression can occasionally lead to outperforming other sophisticated nonlinear models such as ensemble classifier or support vector machine. The formula of logistic regression can be written as:

$$p = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}}, \qquad (4)$$

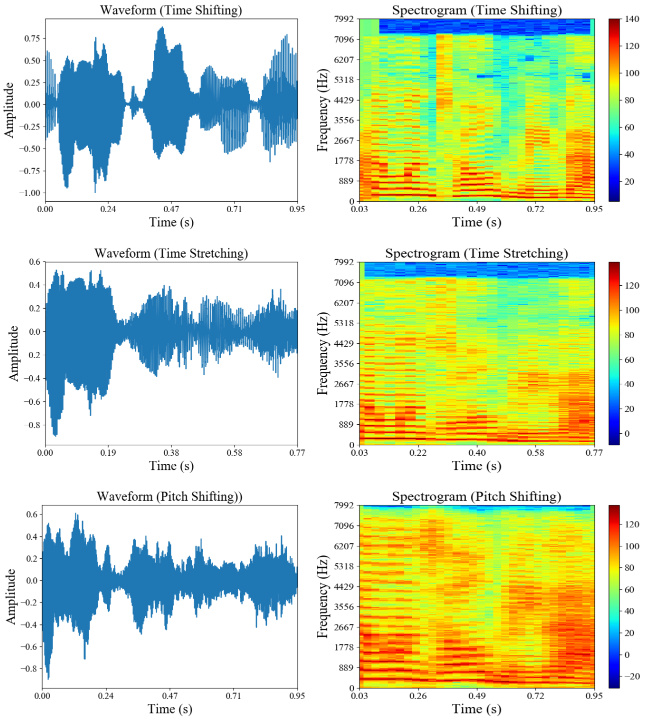

Figure 7: The waveform and spectrogram of the augmented speech clips.

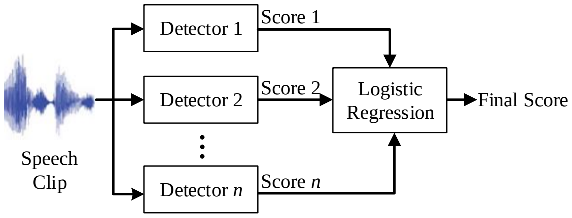

Figure 8: Flowchart of score-level fusion by logistic regression.

where the variables x_i is the output of the ith detector, and parameters β_i for all $i = 0, 1, 2, \ldots, n$ are all estimated. Specifically, we first train detectors on the training dataset and use all detectors' scores to train a logistic regression model. Then, we further calculate the scores of speech clips in testing dataset by all trained detectors. Finally, we send the scores into the trained logistic regression model to produce the final score. The logistic regression is implemented by the Bosaris toolkit[1].

4 DATASETS AND EVALUATION MEASURES

In this section, we describe two datasets used for validating the effectiveness of fake speech detection approaches, and then introduce the metrics to evaluate the performance.

[1] https://sites.google.com/site/bosaristoolkit/

4.1 ASVspoof 2019 LA Dataset

The automatic speaker verification spoof (ASVspoof) challenge is created to foster research on anti-spoofing and to provide common platforms for the assessment and comparison of spoofing countermeasures [32]. The organizers of ASVspoof challenge have published three editions of databases, the last edition of ASVspoof 2019 database varied coverage of different types of spoofing data, could further foster research on anti-spoofing [29]. The ASVspoof 2019 LA dataset is the logical access part of ASVspoof 2019 database, which derived from the VCTK corpus. The ASVspoof 2019 LA dataset is created using speeches from 78 speakers (33 males, 45 females) and stored in 16 kHz at 16 bits-per-sample. The fake speeches are generated from 17 TTS and VC techniques. The entire ASVspoof 2019 LA dataset is partitioned into training, development and evaluation sets. The train and development sets include fake speeches generated from the same speech synthesis techniques (2 VC and 4 TTS techniques). However, only two known techniques are present in the evaluation set and remaining fake speeches are generated from 11 unknown algorithms. More details about the speeches in ASVspoof 2019 LA dataset are shown in Table 1.

Table 1: Illustration of subsets and speeches in ASVspoof 2019 LA dataset.

Subset	Natural	Fake
Training	2,580	22,800
Development	2,548	22,296
Evaluation	7,355	63,882

4.2 FoR-norm Dataset

The Fake or Real (FoR) database is proposed for studies in speech synthesis and synthetic speech detection [18]. The main difference between the FoR database and previous databases is that the FoR database contains speech clips from state-of-the-art speech synthesis algorithms, i.e. speech clips with naturalness similar to real human speech. The fake speeches are synthesized by the latest methodologies, both open source and commercial, such as : Deep Voice, Amazon AWS Polly, Baidu TTS, Google TTS and Microsoft Azure TTS. In addition, the FoR dataset contains a large number of speech clips that are enough to train complex models such as ResNet without over-fitting. To eliminates bias for machine learning experiments, the FoR database is processed into four different versions: FoR-original, FoR-norm, FoR-2seconds and FoR-rerecorded. The FoR-norm dataset is processed by balancing the data to achieve even distribution between genders (male and female) and classes (fake and natural). Due to the balancing process, the FoR-norm version of FoR dataset contains a total of 69,400 speech clips. To eliminate strange and unusual speeches, so we use the FoR-norm version of FoR database. The entire FoR-norm dataset is also partitioned into training, development and evaluation sets. More details about the speeches in FoR-norm dataset are shown in Table 2.

4.3 Evaluation Measures

To evaluating performance of the different detectors on the datasets, we compute two commonly metrics that are used for evaluation of fake speech detection systems. Equal error rate (EER) is used as the first metric for evaluating the fake speech detection methods. The EER is determined by the point at which false acceptance rate and false rejection rate are equal to each other. Mathematically, suppose $P_{\text{fa}}(\theta)$ and $P_{\text{fr}}(\theta)$ stand for the false acceptance rate and false rejection rate at threshold θ respectively, which are defined as :

$$P_{\text{fa}}(\theta) = \frac{\text{Num}\{\text{SFS} > \theta\}}{\text{Num}\{\text{fake speech clips}\}}, \tag{5}$$

$$P_{\text{fr}}(\theta) = \frac{\text{Num}\{\text{SNS} \leq \theta\}}{\text{Num}\{\text{natural speech clips}\}}, \tag{6}$$

where SFS and SNS represent scores of fake speech and scores of natural speech respectively, and Num$\{\cdot\}$ denotes the number of speech clips in the set. The EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal i.e. EER = $P_{\text{fa}}(\theta_{EER}) = P_{\text{fr}}(\theta_{EER})$. After repeating ten times via randomly splitting the training and testing data, we use the average of ten testing errors EER to quantify performance of the different fake speech detection systems. The lower the EER, the more efficient the fake speech detection system.

Table 2: Illustration of subsets and speeches in FoR-norm dataset.

Subset	Natural	Fake
Training	26,941	26,927
Development	5,400	5,398
Evaluation	2,264	2,370

In order to view the performance of detectors for detecting the fake speech from another perspective, we also use receiver operating characteristic (ROC) curve [6] as a secondary metric. The ROC curve illustrates relationship between false positive rate and true positive rate as the discrimination threshold is varied. We draw ROC curve and calculate the area under the ROC curve (AUC) for the results of detection. The fake speech detection systems are extremely vulnerable for fake speeches when the ROC points tend towards the diagonal line, and vice versa.

5 EXPERIMENTS

In this section, we evaluate the performance of the proposed TE-ResNet and several state-of-the-art fake speech detection methods [1, 12, 14, 19, 27] on two public datasets. As we mainly focus on back-end classifiers rather than front-end acoustic features, we have chosen classifiers that used in these related works and named CNN [14], LCNN [12] and ResNet [1]. These back-end classifiers have good feature learning ability and provide competitive performance in general. The experiments are carried on the ASVspoof 2019 LA dataset and FoR-norm dataset, and cross-dataset evaluations are also performed.

5.1 Implementation Details

As supervised learning, the instance of fake speech class is labeled for +1 as positive and the instance of natural speech class is labeled for 0 as negative. The classifiers are expected to calculate the score that the speech clip is fake. During the training phase of neural networks, the stochastic gradient descend (SGD) optimizer Adamax

Table 3: The average EER (%) of detectors on ASVspoof 2019 LA dataset.

Back-end Classifier	Front-end Features	Test Dataset	
		Development	Evaluation
Baseline [27]	-	0.43	9.57
CNN [14]	LPS	0.46	9.36
	MFCC	0.85	8.25
	CQCC	0.65	10.95
LCNN [12]	LPS	0.16	9.12
	MFCC	0.73	7.61
	CQCC	0.66	9.91
ResNet [1]	LPS	0.13	8.68
	MFCC	0.64	7.60
	CQCC	0.63	8.87
TE-ResNet	LPS	0.11	6.02
	MFCC	0.21	6.54
	CQCC	0.19	7.14
Fusion	-	0.10	5.89

[10] is used with mini-batches of 64 speech clips and the training dataset is shuffled after each epoch. The batch normalization parameters are learned via an exponential moving average with decay rate 0.99. At the beginning of training, the filter weights are initialized with random numbers generated from a zero mean truncated Gaussian distribution with standard deviation of 0.1. For the fully connected classifier layers, we initialize the weights with a zero mean Gaussian and standard deviation 0.01 and no bias. The training of all neural networks is run for 500 epochs with an initial learning rate of r_1 = 0.0001. The snapshot model that achieves the best validation performance is taken as the trained neural network model.

The above training strategies are applied for all back-end neural network models. After choosing the trained neural network model, we make the test with mini-batches of 64 speech clips which are randomly sampled from the test dataset without replacement until all data are recycled, and calculate the score of each speech clip. Then the average EER and AUC are calculated according scores of speech clips.

5.2 Experiments on ASVspoof 2019 LA Dataset

In this part, we would like to evaluate the performance of TE-ResNet on ASVspoof 2019 LA dataset. As the ASVspoof 2019 LA dataset is partitioned into training, development and evaluation sets, we evaluate the detectors on development dataset and evaluation dataset respectively. The detection results on the evaluation dataset are calculated to evaluate the performance of the detection methods for unseen fake speech techniques. To give a more intuitive impression of the performance of TE-ResNet, we compare it with some other major fake speech detection approaches which include CNN [14], LCNN [12] and ResNet [1]. The detection results of each approach with front-end acoustic features LPS, MFCC and CQCC are shown in Table 3. In addition, results of the baseline system of ASVspoof 2019 competition [27] and the fusion model are also presented in the Table 3. The first two columns in Table 3 are back-end classifiers and front-end features. The last two columns are the EERs of fake speech detectors on development dataset and evaluation dataset respectively. The last row in Table 3 is the score-level fusion of detection results, which combines the above 12 fake speech detectors (not including baseline model) by logistic regression.

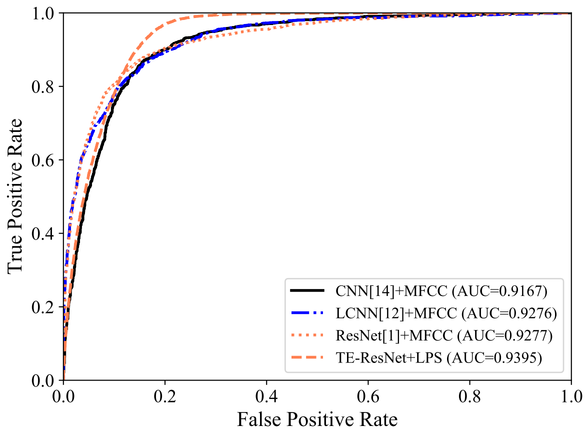

Figure 9: ROC curves of detectors on evaluation set of ASVspoof 2019 LA dataset.

As we can see from Table 3, the EERs of all detectors on evaluation dataset are obviously higher than that on development dataset, which means that the unseen fake speech techniques have a great impact on the detection result and highlights the difficulty of generalizing a fake speech detection system to unseen attack algorithms. The EERs of single detectors on development dataset are all under 0.85% and EERs of single detectors on the evaluation dataset are between 6.02% and 10.95%. For each front-end feature, the TE-ResNet achieves lower EERs than that of other detectors. The best detection results are achieved by TE-ResNet with LPS, which are EERs of 0.11% and 6.02% on development dataset and evaluation dataset respectively. The ResNet [1] achieves pretty lower EERs and has more stable performance over different input acoustic features than CNN [14] and LCNN [12]. The average EER of CNN [14] is the highest and equals to 9.52% on the evaluation dataset, which might due to the vanishing gradient problem. In addition, the proposed score-level fusion system that combines 12 detectors expect for the baseline method, achieves EERs of 0.10% and 5.89% on the development and evaluation dataset respectively, which are approximately a 22.50% relative improvement over the best existing single detection system. Overall, the proposed single fake speech detection scheme has better results than the related existing detectors and the score-level fusion system also has improvement in the detection results.

In order to view the performance of the proposed system for detecting fake speech from another perspective, we draw ROC curve and calculate AUC of fake speech detectors on the evaluation dataset. We choose the best detection result for each back-end classifier, hence get four ROC curves. Figure 9 depicts ROC curves of four fake speech detection methods and the corresponding AUC values are also reported.

As can be seen from Figure 9, the detector based on CNN [14] with MFCC generates the smallest area under ROC curve and the AUC value is 0.9167, which means that the detector is not better

Table 4: The average EER (%) of detectors on FoR-normal dataset.

Back-end Classifier	Front-end Features	Test Dataset Development	 Evaluation
Baseline [19]	-	0.41	6.13
CNN [14]	LPS	0.09	10.25
	MFCC	0.35	10.03
	CQCC	0.33	11.61
LCNN [12]	LPS	0.04	10.68
	MFCC	0.41	7.29
	CQCC	0.26	10.42
ResNet [1]	LPS	0.06	6.05
	MFCC	0.17	5.01
	CQCC	0.22	6.09
TE-ResNet	LPS	0.03	4.38
	MFCC	0.12	6.08
	CQCC	0.16	5.77
Fusion	-	0.02	3.99

when detects fake speech. The AUC of the TE-ResNet with LPS is higher than the related detectors. The results from ROC curves also verify that the TE-ResNet provides better balance between values of false positive and true positive, which makes it more practical for detecting the fake speech.

5.3 Experiments on FoR-normal Dataset

In this part, we would like to evaluate the performance of TE-ResNet on the FoR-normal dataset. The comparative detection methods are CNN [14], LCNN [12] and ResNet [1]. In addition, results of the baseline system [19] on FoR-normal dataset are reported. To evaluate the detection methods on unseen fake speech techniques, we calculate the detection score on the evaluation dataset. The values of EER for different fake speech detection approaches on FoR-normal dataset are described in Table 4.

From Table 4, it can be seen that the EERs of all detectors on evaluation dataset are obviously higher than that on development dataset, which is same conclusion as that on ASVspoof 2019 LA dataset. This verifies that it is difficult to detect the unseen fake speech techniques. On the other hand, it can be seen that the proposed TE-ResNet with LPS feature outperforms the related methods, and achieves the EERs of 0.03% and 4.38% on the development and evaluation dataset respectively, which are approximately a 12.57% improvement over the best existing single detection systems. The score-level fusion system, which is implemented as the same as in Section 5.2, gives EERs of 0.02% and 3.99% on the development and evaluation dataset respectively. In addition, the overall detection result on the FoR-normal dataset is batter than that on the ASVspoof 2019 LA dataset, which might indicate that diversity of fake speech techniques has a heavily influence for the fake speech detection systems.

Furthermore, we also draw ROC curve and calculate AUC for fake speech detectors on the evaluation dataset in Figure 10. As can be seen from Figure 10, the detector that based on CNN [14] and MFCC generates the smallest area under ROC curve and the AUC value is 0.8913. The AUCs of LCNN [12] and ResNet [1] methods are higher than that of CNN [14]. The AUC of TE-ResNet with LPS equals to 0.9535, which has higher AUC than that of the existing detectors.

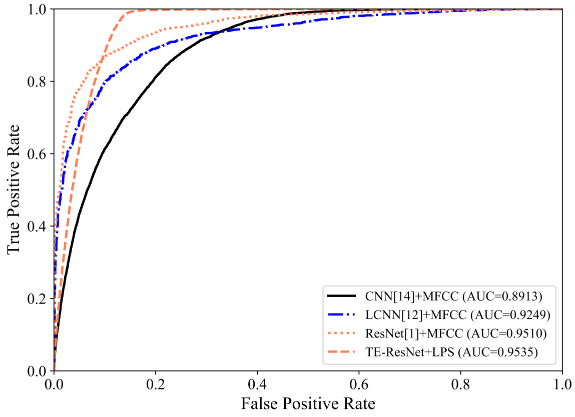

Figure 10: ROC curves of detectors on evaluation set of FoR-normal dataset.

Table 5: The average EER (%) of detectors for cross-dataset evaluation.

Back-end Classifier	Front-end Features	Test Dataset ASVspoof 2019 LA	 FoR-normal
CNN [14]	LPS	28.09	28.12
	MFCC	23.33	20.28
	CQCC	24.35	22.22
LCNN [12]	LPS	23.77	24.04
	MFCC	21.06	19.46
	CQCC	25.87	22.22
ResNet [1]	LPS	19.56	20.12
	MFCC	20.31	18.58
	CQCC	22.98	22.56
TE-ResNet	LPS	19.23	18.15
	MFCC	20.16	19.12
	CQCC	20.98	21.01
Fusion	-	19.02	18.13

5.4 Cross-dataset Evaluations

Given two labeled datasets that target for fake speech detection, cross-dataset evaluations aim to detect the fake speeches that generated by completely unseen fake speech technologies. To explore how well the detection model trained from the one dataset generalizes to other dataset, we train the detectors on the ASVspoof 2019 LA dataset and test it on FoR-normal dataset, and vice versa. To compared the performance of TE-ResNet with the existing state-of-the-art fake speech detection methods, we also make cross-database evaluations for CNN [14], LCNN [12] and ResNet [1]. The values of EER for different fake speech detection systems are described in Table 5. The last column in Table 5 is EERs of fake speech detectors on the entire FoR-normal dataset, which are trained on the entire ASVspoof 2019 LA dataset.

From Table 5, we can see that the overall EERs of all detectors are higher than 18.15%. For evaluation on ASVspoof 2019 LA dataset,

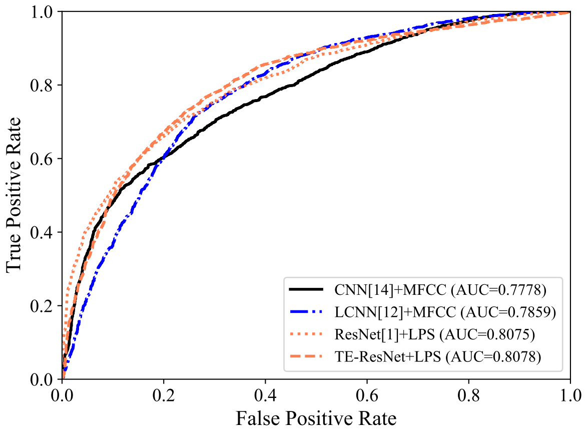

Figure 11: ROC curves of cross-dataset evaluation (training on FoR-normal dataset and evaluation on ASVspoof 2019 LA dataset).

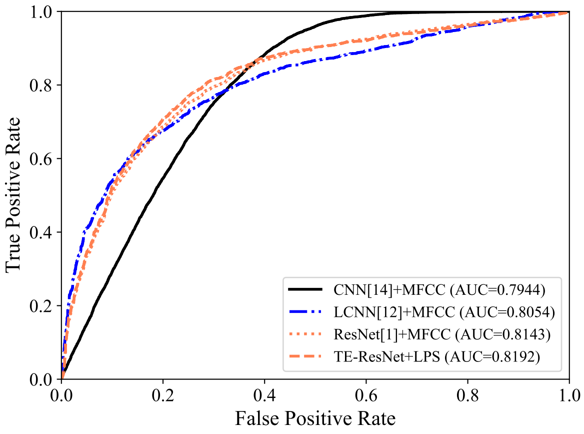

Figure 12: ROC curves of cross-dataset evaluation (training on ASVspoof 2019 LA dataset and evaluation on FoR-normal dataset).

the best performance among the existing fake speech detectors is achieved by ResNet [1] with LPS, which gives an EER equal to 19.56%. The TE-ResNet with LPS achieves an EER of 19.23% when detect speech clips on ASVspoof 2019 LA dataset. For evaluation on FoR-normal dataset, the lowest EER of 18.15% is achieved by TE-ResNet with LPS and the CNN [14] with LPS obtain the highest EER of 28.12%. In addition, comparing the last two columns of the Table 5, we can see that the EERs of all detectors on ASVspoof 2019 LA dataset are higher than that on the FoR-normal dataset. It shows that when we use the speech clips, which are generated by complex and diverse fake speech techniques, to train the detectors, the trained detector can be performed well on the fake speeches that generated by a few fake speech techniques. In a word, the existing fake speech detectors are not well for cross-database evaluations and there is a large room to improvement.

To show the performance of fake speech detectors with varying thresholds, we draw ROC curve and calculate AUC for each backend classifier with the best front-end feature on ASVspoof 2019 LA dataset and FoR-normal dataset respectively. The ROC curves of different fake speech detection methods are depicted in Figure 11 and Figure 12, and the corresponding AUC values are also reported.

Comparing Figure 11 and Figure 12, the ROC curves in Figure 12 is more close to the top left than the ROC curves in Figure 11. The proposed TE-ResNet with LPS feature generates the largest area under ROC curve on both datasets. The AUC values of the TE-ResNet with LPS feature are 0.8078 and 0.8192 on ASVspoof 2019 LA dataset and FoR-normal dataset respectively. The results from ROC curves also verify that all existing fake speech detectors are not well for cross-database evaluations especially for the increasingly diversity of fake speech techniques.

6 CONCLUSION

In this paper, we propose a new fake speech detection scheme to improve the detection performance, which mainly consists of three parts: speech data augmentation, residual network with transformer encoder, and score-level fusion. By introducing five traditional speech data augmentation strategies, the quantity of training data is increased by five times. TE-ResNet is used to pre-process the acoustic features and calculate score that the speech is fake. With score-level fusion of logistic regression, the fused model takes full advantages of the 12 single detection models. Experimental results on ASVspoof 2019 LA dataset and FoR-normal dataset prove the effectiveness of the proposed TE-ResNet for fake speech detection task, and significantly improves the detection performance compared with other major state-of-the-art approaches. In the future, we will explore more efficient methods to enhance the robustness of fake speech detection.

ACKNOWLEDGMENTS

This work was supported by NSFC under 61902391, 61972390, U1736214, 61802393 and 61872356, National Key Technology R&D Program under 2019QY2202, 2019QY(Y)0207 and 2020AAA0140000, and IIE CAS Climbing Program.

REFERENCES

[1] Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. 2019. Deep residual neural networks for audio spoofing detection. *Proc. Interspeech 2019* (2019), 1078–1082.
[2] Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, and Kai Yu. 2015. Robust deep feature for spoofing detection—The SJTU system for ASVspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association.*
[3] Bhusan Chettri, Daniel Stoller, Veronica Morfi, Marco A Martínez Ramírez, Emmanouil Benetos, and Bob L Sturm. 2019. Ensemble models for spoofing detection in automatic speaker verification. *arXiv preprint arXiv:1904.04589* (2019).
[4] Rohan Kumar Das, Xiaohai Tian, Tomi Kinnunen, and Haizhou Li. 2020. The attacker's perspective on automatic speaker verification: An overview. *Proc. Interspeech 2020* (2020), 4213–4217.
[5] Zijun Deng, Lei Zhu, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Qing Zhang, Jing Qin, and Pheng-Ann Heng. 2019. Deep multi-model fusion for single-image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2453–2462.
[6] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[7] Jee-weon Jung, Hye-jin Shim, Hee-Soo Heo, and Ha-Jin Yu. 2019. Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge. *arXiv preprint arXiv:1904.10134* (2019).

[8] Madhu R Kamble, Hardik B Sailor, Hemant A Patil, and Haizhou Li. 2020. Advances in anti-spoofing: From the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing* 9 (2020).

[9] KL Kashima and B Mont-Reynaud. 1985. The bounded-Q approach to time-varying spectral analysis. *Dept. of Music, Stanford Univ., Tech. Rep. STAN-M-28* (1985).

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. 2017. Audio replay attack detection with deep learning frameworks. *Proc. Interspeech 2017* (2017), 82–86.

[12] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. STC antispoofing systems for the aSVspoof2019 challenge. *Proc. Interspeech 2019* (2019), 1033–1037.

[13] Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7689–7693.

[14] Tijana Nosek, Siniša Suzić, Boris Papić, and Nikša Jakovljević. 2019. Synthesized speech detection based on spectrogram and convolutional neural networks. In *2019 27th Telecommunications Forum (TELFOR)*. IEEE, 1–4.

[15] Alan V Oppenheim, John R Buck, and Ronald W Schafer. 2001. *Discrete-time signal processing.* Upper Saddle River, NJ: Prentice Hall.

[16] Monisankha Pal, Dipjyoti Paul, and Goutam Saha. 2018. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language* 48 (2018), 31–50.

[17] Dipjyoti Paul, Monisankha Pal, and Goutam Saha. 2017. Spectral features for synthetic speech detection. *IEEE journal of selected topics in signal processing* 11, 4 (2017), 605–617.

[18] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 1–10.

[19] Ricardo Amaral Martins Reimao. 2019. Synthetic speech detection using deep neural networks. (2019).

[20] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. In *Sixteenth Annual Conference of the International Speech Communication Association.*

[21] Suleiman Usman Santuraki. 2019. Trends in the regulation of hate speech and fake news: A threat to free speech? *Hasanuddin Law Review* 5, 2 (2019), 140–158.

[22] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel. 2015. Joint speaker verification and antispoofing in the *i*-vector space. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 821–832.

[23] Anagha Sonawane and MU Inamdar. 2017. Synthetic speech spoofing detection using MFCC and SVM. *International Journal of Advance Research, Ideas and Innovations in Technology* (2017).

[24] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco. 2020. Spoofing attack detection using the non-linear fusion of sub-band classifiers. *Proc. Interspeech 2020* (2020), 1106–1110.

[25] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45 (2017), 516–535.

[26] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. 2016. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients.. In *Odyssey*, Vol. 2016. 283–290.

[27] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H Kinnunen, and Kong Aik Lee. 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. *Proc. Interspeech 2019* (2019), 1008–1012.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 6000–6010.

[29] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language* 64 (2020), 101114.

[30] Shengyun Wei, Shun Zou, Feifan Liao, et al. 2020. A comparison on data augmentation methods based on deep learning for audio classification. In *Journal of Physics: Conference Series*, Vol. 1453. IOP Publishing, 012085.

[31] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah. 2012. A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–5.

[32] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Héctor Delgado. 2017. ASVspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing* 11, 4 (2017), 588–604.