Detection of Voice Conversion Spoofing Attacks using Voiced Speech

Arun Sankar M. S.[1], Phillip L. De Leon[2], and Utz Roedig[1]

[1] School of Computer Science and Information Technology, Cork, Ireland
asankar@ucc.ie,u.roedig@ucc.ie
[2] New Mexico State University, Klipsch School of Electrical and Computer Engineering, Las Cruces, New Mexico, U.S.A.
pdeleon@nmsu.edu

Abstract. Speech consists of voiced and unvoiced segments that differ in their production process and exhibit different characteristics. In this paper, we investigate the spectral differences between bonafide and spoofed speech for voiced and unvoiced speech segments. We observe that the largest spectral differences lie in the 0-4 kHz band of voiced speech. Based on this observation, we propose a low-complexity, pre-processing stage which subsamples voiced frames prior to spoofing detection. The proposed pre-processing stage is applied to two systems, LFCC+GMM and IA/IF+KNN that differ entirely on the features and classifier used for spoofing detection. Our results show improvement with both systems in detection of the ASVspoof 2019 A17 voice conversion attack, which is recognized to have one of the highest spoofing capabilities. We also show improvements in the A18 and A19 voice conversion attacks for the IA/IF+KNN system. The resulting A17 EERs are lower than all reported systems where the A17 spoofing attack is the worst attack except the Capsule Network. Finally, we note that the proposed pre-processing stage reduces the speech date by more than 4× due to subsampling and using only voiced frames but at the same time maintaining similar pooled EER as that for the baseline systems, which may be advantageous for resource constrained spoofing detectors.

Keywords: Spoofing detection · Speech processing· Computer security· Voice bio-metric

1 Introduction

Traditionally, usernames and passwords are used for authentication. However, handling usernames and passwords securely has been proven to be difficult and compromised passwords have lead to many security breaches. The burden of using passwords can be eliminated by using biometric authentication. For example, finger prints, retina scans or voice prints can be used as input for authentication.

Automatic Speaker Verification (ASV) systems are popular as a low-cost and flexible technology for biometric authentication. However, even these systems

are known to be vulnerable to spoofing which can be classified into attacks via impersonation, replay, speech synthesis, twins, and voice conversion [1]. Among these, replay, speech synthesis, and voice conversion remain threats due to the availability of successful open-source tools for generating high-quality spoofed speech which can be used in a targeted attack [2].

Countermeasures to detect spoofed speech and thus prevent an attack, are in active development and the ASVspoof challenge, initiated in 2015, has assisted with advancing the research through organized trials and evaluations [3]. Most developed methods perform feature extraction in the frequency domain using filter banks to obtain sub-band spectral features. The features are analysed using sophisticated classifiers such as Gaussian Mixture Model (GMM) or Deep Neural Networks (DNN), and the best performing systems use a number of classifiers in combination, i.e. ensemble classifier. There have been significant advances in spoofing detection to the point where top-performing systems evaluated using the ASVspoof 2019 dataset report pooled min-tandem-Decision Cost Function (t-DCF) below 0.1 and Equal Error Rate (EER) below 3.5% (see [4]). Recently 12 state-of-the-art detection systems have been reported in [5] and evaluated using the ASVspoof 2019 dataset. It was found that the most successful spoofing attacks are A08 (most successful for 2 systems), A17 (most successful for 9 systems), and A18 (most successful for 1 system). Attack A08 is speech synthesis and attacks A17 and A18 are voice conversion. However, state-of-the-art systems' performance against the worst ASVspoof 2019 attacks have an average EER of 12.94% [5]. Of the 9 systems reporting A17 as the worst attack, the average EER is 14.2% with Capsule Network reporting 3.76% EER [5]. Thus for some specific attacks, detection accuracy is still lacking.

The speech signal is composed of voiced and unvoiced segments that differ by the production mechanism and characteristic features [6]. These segments are separately used for many speech processing applications due to the difference in the type and depth of information contained in these segments. For example, the speaker-specific unique information can be found much in voiced segments due to vocal cord vibration and so on [7]. In general, spoofing attacks are applied to the entire speech signal without considering separately voiced and unvoiced segments and hence the location and level of artefacts vary with these segments.

In this paper, we investigate the spectral differences between human (bonafide) and spoofed speech for voiced and unvoiced speech segments. When comparing spectra of bonafide and spoofed speech, we find the largest differences lie in voiced segments in the 0-4 kHz band. With this observation, we propose a low-complexity pre-processing stage which *subsamples voiced frames* prior to spoofing detection. We evaluate this novel pre-processing stage using different detection systems. The core contribution of this work is the insight that voiced and unvoiced speech segments contribute very differently to the task of spoofing detection.

Our specific contributions are as follows:

– We show that voiced speech segments are more useful for spoofing detection than unvoiced speech segments. We also describe the distribution of infor-

mation for spoofing detection over frequency bands in voiced and unvoiced speech segments.
- We propose a low-complexity pre-processing stage which subsamples only voiced frames prior to spoofing detection. This pre-processing stage reduces the amount of necessary data by a factor of 4 while maintaining overall detection accuracy (similar pooled EER).
- We show that this pre-processing stage can be combined with different existing spoofing detection systems.
- We show an improvement in the detection accuracy for the challenging ASVspoof 2019 A17 voice conversion attack using two different detection systems together with the novel pre-processing stage. We also show improvements for the A18 and A19 voice conversion attacks in some settings.

This paper is organized as follows. The details of the ASVspoof database used for conducting experiments are given in Section 2. In Section 3, we provide a brief review of speech production focusing on voiced speech and place of articulation as motivation for the investigation of using voiced speech for spoofing detection. In Section 4, we present our observations on the spectral differences between bonafide and spoofed speech for voiced and unvoiced segments. In Section 5, we propose a pre-processing stage which takes as input the speech signal and passes to the countermeasure a signal containing only voiced segments and in Section 6 we provide detection results for two different countermeasures with and without the pre-processing stage. Section 7 summarizes the works done in spoofing detection and how our work differs from others. In Section 8, we discuss the results paying close attention to the A17 attack which is considered the most difficult attack to detect. Finally, in Section 9, we conclude the paper.

2 ASVspoof Challenge Dataset and Evaluation Metric

The ASVspoof challenge series was initiated in 2015 with the motivation of advancing spoofing detection and countermeasures. The first challenge was focused on voice conversion and synthetic speech attacks while the second spoof challenge organized in 2017 concentrated on replay attacks as they are much easier to generate without any technical expertise. The third spoof challenge took place in 2019 and considered speech synthesis, voice conversion, and replay attacks. The fourth challenge organized recently in 2021 focused on discriminating between genuine and spoofed or deepfake speech using ASVspoof 2019 database.

The ASVspoof 2019 challenge database consists of a logical access (LA) partition containing voice conversion and speech synthesis examples in addition to the physical access (PA) partition which contains replay examples. Each partition contains training, development and evaluation subsets. The training and development subsets are used for conducting experiments related to the development of the detection model while the evaluation set is utilized for measuring detection performance of the developed model. The training and development subsets of LA contain 6 spoofing attacks which are considered as known attacks and used for the construction of the detection model. The evaluation subset of

LA has 11 unknown attacks to determine the efficiency of the developed model on attacks that are unknown to the system or in other words on attacks that are not used for training the model. In addition, each subset also contains examples of human-produced speech. All speech examples, including the source utterances for creating the spoofed speech, are taken from the VCTK corpus [8]. The utterances consist of 107 speakers (46 male and 61 female) that are partitioned into three disjoint subsets. Details of the database are summarized in Table 1.

Table 1. Description of the logical access partition of the ASVspoof 2019 challenge database.

Database attributes	**Training set**	**Development set**	**Evaluation set**
Spoofing attack algorithms	A01-A06	A01-A06	A07-A19
Spoofing methods	TTS (4) VC (2)	TTS (4) VC (2)	TTS (6) VC (2) Hybrid (3)
Known attacks	6	6	2 (A16=A04, A19=A06)
Unknown attacks	0	0	11
No. of genuine samples	2580	2548	7355
No. of spoofed samples	22800 (3800×6)	22296 (3716×6)	63882 (10647×6)
No. of male speakers	8	4	21
No. of female speakers	12	6	27

The training and development data sets are built using the same set of spoofing attacks (A01-A06). Spoofing attacks A01 to A04 are based on Text-to-Speech (TTS) methods while attacks A05 and A06 use voice conversion (VC) methods. The attacks A01-A03 are neural network based TTS systems and attack A04 does TTS using waveform concatenation method. The evaluation data set consists of 13 spoofing attacks (A07-A19) out of which 2 attacks (A16 and A19) are considered as known attacks and the remaining 11 spoofing attacks are unknown attacks. Attacks A16 and A19 use the same spoofing techniques as attacks A04 and A06 respectively. The unknown attacks consist of six TTS based methods (A07-A12), two VC methods (A17 and A18) and three hybrid models (A13-A15). The hybrid models use a combination of VC and TTS for the generation of spoofed speech.

The following metrics are used for quantifying the detection performance of spoofing detector.

– *Equal Error Rate (EER)* - An ideal spoofing detector should flag spoofed speech and pass genuine speech but in reality there is always some error which is quantified using False Acceptance Rate (FAR) and False Rejection Rate (FRR).
False Acceptance Rate: It is the ratio of spoofed speech samples wrongly classified as genuine speech and can be written as

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{1}$$

where False Positive (FP) is the number of spoofed speech samples misclassified as genuine speech and True Negative (TN) denotes the number of correctly identified spoofed speech samples.
False Rejection Rate: It is defined as the ratio of genuine samples misclassified as spoofed speech. FRR can be expressed as

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{2}$$

where True Positive (TP) is the correctly identified bonafide speech samples and False Negative (FN) is the number of genuine speech samples misclassified as spoofed speech.
It is desirable to minimize both FAR and FRR for improving the efficiency of detection systems. But adjusting the detection threshold to reduce either of the errors harm the other. The detection threshold plot has a point where both the error rates are equal and that common value is called the EER which is considered a metric in ASV spoof 2019 challenge.
– *tandem-Decision Cost Function:* The EER metric is sufficient to quantify the performance of a stand alone spoofing detector. But when this detector is integrated into an ASV system, the impact of countermeasure on verification performance cannot be evaluated by EER metric. In such scenario, the t-DCF metric [9] measures the impact of spoofing and countermeasure on the reliability of ASV system by combining the verification and spoofing errors. The minimum normalized tandem-Decision Cost Function is expressed in the form

$$\text{t-DCF}_{\min} = \min_{Thr} \left\{ \beta P_{cm\,MISS}(Thr) + P_{cm\,FAR}(Thr) \right\}. \tag{3}$$

The parameter β depends on the spoofing prior and cost parameters and on miss and false alarm rates of speaker verification. $P_{cm\,MISS}(Thr)$ and $P_{cm\,FAR}(Thr)$ are the false alarm and miss rates of the counter measure at threshold Thr.

For additional information, please see [10].

3 Brief Review of Speech Production

Speech is an acoustic wave produced by the air expelled from lungs which serves as the excitation for the acoustic filter consisting of vocal and nasal tracts [6].

The frequency spectrum of the excitation is shaped by the frequency selectivity of these tracts. The vocal tract contains different sections called articulators that play a crucial role in the generation of different sounds by shaping the airflow. During speech production, the airflow is modulated according to the sound to be generated by the movement of active articulators toward the passive articulators which remain stationary throughout the process [11]. This relative placement of articulators will create different types of constrictions for generating various voiced (vowels) and unvoiced (plosives, consonants) sounds. The features of vocal and nasal tracts change continuously with time and make speech radiated from lips non-stationary.

The basic difference between voiced and unvoiced sound is due to the behavior of vocal cords during sound production [6]. During vocal cord vibration, air flowing from the lungs will be interrupted periodically by the vocal cords providing a series of pulses for excitation of the vocal tract which produces voiced speech signals. Voiced speech is dominated by periodic pulses and a set of formants which are peaks in the frequency spectrum due to the acoustic resonance of vocal tract. These spectral peaks are in the low-frequency region and hence the energy of voiced speech mostly lies below 4 kHz. When vocal cords remain stationary, the vocal tract will have a random excitation and constriction by different articulators and will generate unvoiced sound. These unvoiced sounds are non-periodic, sounds random and their energy is mainly contained in region from 2-8 kHz [12]. The speech production mechanism is thus modelled as a source excitation passing through a time-varying filter that corresponds to the dynamic characteristics of vocal tract. The excitation is random noise for unvoiced sounds and a series of pulses for voiced sounds which represents the fundamental frequency in speech.

Spoofed speech is generated using TTS and VC techniques in order to change the voice identity of speech to that of a target speaker to be perceived true by humans and/or speaker verification systems. In ASVspoof 2019 challenge database, the spoofed speech is generated using four TTS (A01-A04) and two VC (A05-A06) spoofing attack algorithms for training and development sets and by using ten TTS (A07-A16) and three VC (A17-A19) spoofing attack algorithms for the evaluation set [13].

The spoofed speech generation using the TTS system converts the input text to speech that feels like to be spoken by the target speaker. This process involves conversion of text to linguistic features and then to acoustic features which are used to generate the waveform of desired speech. The VC techniques change the voice identity of speech without changing the linguistic content. When parallel training data (utterances with the same linguistic content for both source and target speakers) is available, the VC can be easily done using methods such as dynamic time wrapping and spectral mapping [14, 15]. Due to the difficulty in obtaining parallel training data, many VC methods are developed using non-parallel training data. In all these methods, the input speech undergoes an intermediate transformation to remove the source speaker characteristics followed by the addition of target speaker characteristics and reconstruction of speech.

The various VC models include variational auto-encoder (A05 and A17), GMM-UBM with speech source filter model (A06), i-vector Probabilistic Linear Discriminant Analysis (PLDA) based transfer learning, and so on [13, 16]. The spoofing attack A06 does VC for the generation of spoofed speech by mapping the source-filter characteristics of input speech on a frame-by-frame basis to that of the target speaker. The input audio signal is analyzed and the derived acoustic features (Mel frequency Cepstrum Coefficients (MFCC) and Linear Prediction Cepstrum Coefficients (LPCC)) are modified to match the filter characteristics with that of the target speaker and in order to produce the spoofed speech. The spoofing attack A18 uses a transfer learning method to predict the i-vectors of target speaker from the i-vectors of source speaker. The knowledge about predicted i-vectors are used to generate the MFCCs of target speaker and thereby for the production of spoofed speech. The attacks A05 and A17 use variational auto-encoder for mapping the spectral features of input audio from source to target speaker. The auto-encoder is trained to encode the incoming spectral feature vectors to speaker independent vectors and then to decode them with the characteristics for the target speaker. This is followed by a speech reconstruction process in which attacks A05 and A17 differ.

The spoofing attack A17 uses direct waveform modification method for the generation of target speech but the spoofing attack A06 uses WORLD vocoder. In direct waveform modification, spectral details are preserved that help in producing high-quality speech. The target speech waveform is generated in spoofing attack A17 by passing the $F0$ transferred residual signal through a synthesis filter designed for the target speaker using the converted spectral features. The $F0$ transferred residual signal is sensitive to the spectral estimation error due to the difficulty in modelling speaker characteristics which is a problem associated with conversion models based on non-parallel training data. The interaction of $F0$ transformed residual signal with inaccurately estimated spectrum will produce noise in the reconstructed speech. It is evident from the spectral plots of source speech and target speech obtained using waveform modification method given in [17] that spectral errors are prominent in the low-frequency region.

Motivated by the spectral estimation errors and the fact that speech (excluding silence) is dominated by voiced sounds that contain mainly low-frequency components, use of these voiced segments may provide better features for discriminating between genuine and spoofed speech generated using VC methods based on direct waveform modification.

4 Spectral Differences in Voiced/Unvoiced Segments from Human and Spoofed Speech

Initial work in detection of spoofed speech, extracted discriminating features from the entire speech signal and using pre-trained models, classified speech signals as genuine or spoofed [18, 19]. Later work extracted features from specific components of the decomposed speech signal which contained more discriminating information than the signal as a whole, in order to improve detection

accuracy. For example, spoofing detectors based on specific words [20], based on specific spectral bands [21, 22], or based on specific modes in Empirical Mode Decomposition (EMD) have been investigated [23]. Speech signals are generally composed of voiced, unvoiced, and silence segments [11]. Typically, little if any discriminating features exist in silence segments and thus we focus on analyzing unique discriminating features within voiced and unvoiced segments. To the best of our knowledge spoofing detectors based on voiced or unvoiced segments have not been investigated.

In order to analyze spectral differences in voiced and unvoiced segments from genuine and spoofed speech, we choose from ASVspoof 2019 LA training sets A01-A04 (TTS), two male speakers (LA92 and LA95) and two female speakers (LA79 and LA80). Next, we identified identical sentences from genuine and spoofed speech for each speaker. Next, for each identical sentence pair, we segmented phonemes according to voiced or unvoiced. Finally for each voiced/unvoiced segment we computed the difference in magnitude spectra between the genuine and spoofed segment. For training sets A05-A06 (VC), identical sentences did not exist so we identified identical words from genuine and spoofed speech for each speaker and proceeded as above with phoneme segmentation and computation of the difference spectra. The difference spectra were then averaged and are shown in Figs. 1 and 2.

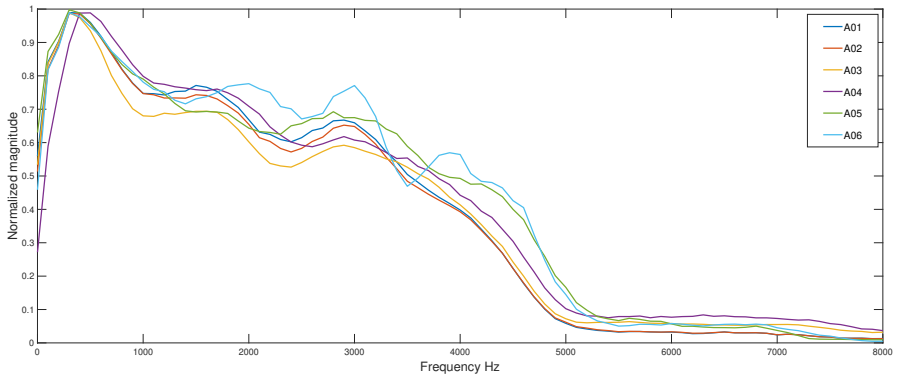

Fig. 1. Average difference (between bonafide and spoofed speech) magnitude spectra for voiced segments from ASVspoof 2019 attacks A01 to A06. We observe a large, well-defined spectral difference in the 0-4 kHz frequency band.

When comparing spectra of human (bonafide) and spoofed speech, we find the largest differences lie in voiced segments over the 0-4 kHz band; smaller differences exist in unvoiced segments over the 4-8 kHz band. From this observation, we find that we are able to accurately classify bonafide versus spoofed

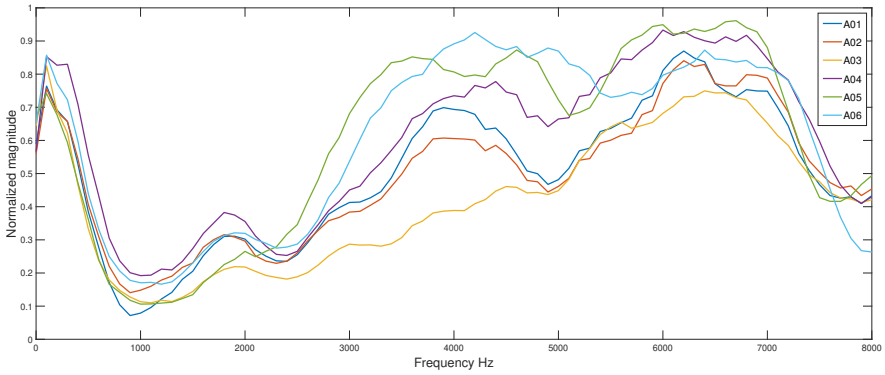

Fig. 2. Average difference (between bonafide and spoofed speech) magnitude spectra for unvoiced segments from ASVspoof 2019 attacks A01 to A06. Unlike voiced speech (Fig. 1), we observe an uneven spectral difference in the unvoiced speech, however, most of the spectral difference lies in 4-8 kHz frequency band.

speech using only voiced speech segments from the 0-4 kHz band. When viewed as a general pre-processing stage, we can show this technique, i.e. using only voiced segments from 0-4 kHz, can be applied to various detectors including Linear frequency Cepstrum Coefficients (LFCC)+GMM while maintaining similar accuracy. By using only voiced segments and downsampling the signal to 4 kHz bandwidth, the data rate and hence computation can be reduced.

From Fig. 1 we observe differences in spectra for voiced segments in the 0-4 kHz band where the largest differences are in the 0-1 kHz band. A simple linear interpolation over 300-4000 Hz shows an approximate difference rate of -17 dB/kHz. On the other hand, in the band from 4-8 kHz the spectral difference is minor. This suggests that for voiced segments, most of the spectral discriminating features lie in 0-4 kHz band. From Fig. 2 we also observe differences in spectra for unvoiced segments but in the 4-8 kHz band. Furthermore, these differences are not as great as in the voiced segments. Given these observations, in the next section we propose a pre-processing stage for which features are extracted from voiced segments only.

5 Subsampling and Voiced Segmentation as a Pre-Processing Stage

From our observations of the spectral differences in voiced segments from bonafide and spoofed speech, we propose a pre-processing stage which takes as input the speech signal and passes to the countermeasure a signal containing only voiced segments. Furthermore, because most of the spectral difference in the voiced segment lies in the 0-4 kHz band, we may subsample the signal by $2\times$. In the

implementation, shown in Fig. 3, we first use 20 ms speech frames and a Zero Crossing Rate (ZCR) detector to label the frames as voiced or unvoiced. We subsample the speech signal by 2× and retain only the corresponding voiced frames. The proposed pre-processing stage lowers the date rate approximately by a factor of 4, i.e. removal of silence and unvoiced segments shortens the signal by approximately half and downsampling reduces the data by another half. This reduction in data may be important in applications where low-complexity spoofing detection is important, e.g. Personal Voice Assistants (PVAs).

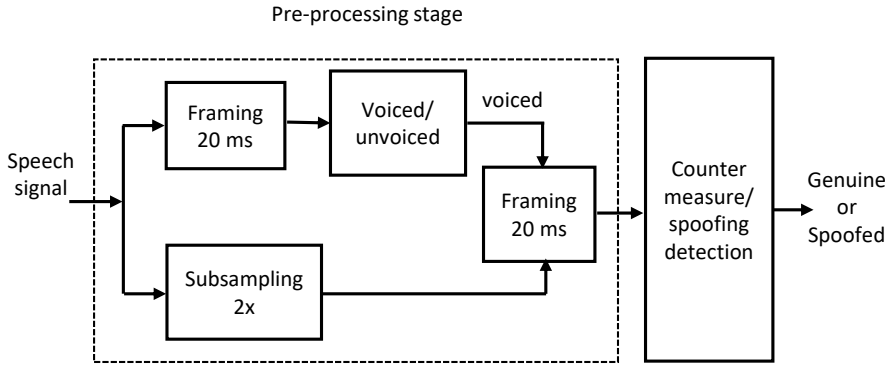

Fig. 3. Block diagram of the proposed pre-processing stage.

6 Spoofing Detection Results using Proposed Pre-Processing Stage

In order to test the proposed pre-processing stage, we consider two detectors which use different features and have results which are among the top performing systems, excluding ensemble systems, using the ASVspoof 2019 evaluation set. We exclude neural networks and deep learning systems in order to focus on the generalization of this approach as a pre-processing stage to conventional systems. The first system uses LFCC features with a ML detector based on a GMM [22] and the second system uses statistics of the Instantaneous Amplitude (IA)/Instantaneous Frequency (IF) from Intrinsic Mode Functions (IMFs) decomposed using EMD [24].

6.1 Brief Overview of Anti-Spoofing Systems used in this Work

The first system under consideration is the LFCC-based system proposed in [22]. The LFCC features are extracted from the entire speech signal spectrum using a filter order of 70. The training and development sets of ASVspoof 2019 challenge

are used to generate the LFCC features and to train the GMM using 1024 components. Classification uses ML estimation. For additional details on this system, we refer the reader to [22]. As software for this system was unavailable, we implemented our own version. Results of this system for ASVspoof 2019 evaluation set A07-A19 and pooled baseline results are given in the second row of Table 2. Our implementation has a slightly higher EER (3.85%) than the published result (3.51%) which may be due to the difference in the environment.

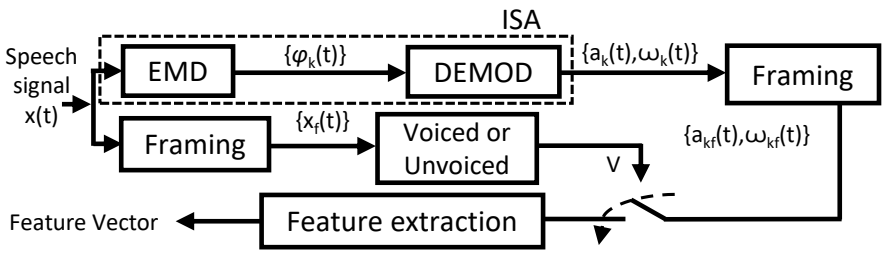

Fig. 4. Block diagram of the EMD based feature vector creation for spoofing detection.

The second system under consideration is based on statistics of the IAs and IFs from IMFs resulting from EMD as proposed in [24]. In this system shown in Fig. 4, EMD is used to decompose the entire speech signal into IMFs. IMFs are each demodulated in order to obtain the IAs and IFs. For each of the first 10 IMFs, we compute the statistics $\{\mu, \sigma^2, \gamma, \kappa\}$ of the IA and IF resulting in an 80×1 feature vector. We use a k-Nearest Neighbours (KNN) classifier to determine whether the speech signal is bonafide or spoofed. For additional details on this system, we refer the reader to [24]. Results of this system for ASVspoof 2019 evalaution set A07-A19 and pooled baseline results are given in the third column of Table 2.

6.2 Results using Proposed Pre-Processing Stage

The proposed pre-processing stage can be viewed as a front-end to the anti-spoofing countermeasure. This front-end subsamples the speech signal by $2\times$ and retains only those frames which have been classified as voiced. Since voiced frames are approximately half of the speech signal (unvoiced and silence frames are the other half), this front-end reduces the data presented to the countermeasure by approximately $4\times$ and hence lowers computation. The LFCC-GMM detector with the proposed pre-processing stage gives a pooled EER and min-t-DCF of respectively 3.99% and 0.10% which is slightly worse compared to the baseline system's EER and roughly the same for the min-t-DCF. While comparing the performance for each individual attack (A7-A19), the proposed methods improves the EER for 3 attacks (A12, A15, and A17); gives worse EER for 5 attacks (A10, A13, A14, A18, and A19); and nearly the same (within 0.05%)

Table 2. The spoofing detection error metrics (%) for LFCC+GMM and IA/IF-KNN systems using only voiced segments or the entire speech signal (baseline). Error rates in red denote worse performance when using voiced segments, while those in blue denote better performance when using voiced segments.

Spoofing attack algorithms	**Spoofing Detectors**			
	LFCC-GMM (Baseline)	**LFCC-GMM (Voiced)**	**IAIF-KNN (Baseline)**	**IAIF-KNN (Voiced)**
A07	0.02	0.02	2.09	3.50
A08	0.00	0.02	2.09	3.50
A09	0.00	0.02	2.09	3.50
A10	12.74	14.89	2.09	3.50
A11	0.00	0.02	2.09	3.50
A12	1.87	1.77	2.09	3.50
A13	2.87	3.44	2.09	3.50
A14	0.00	0.76	2.09	3.50
A15	2.01	1.63	2.09	3.50
A16	0.02	0.02	2.09	3.50
A17	7.47	3.97	11.90	4.56
A18	0.04	2.73	10.15	4.46
A19	0.08	0.14	6.63	3.50
Pooled tDCF	0.10	0.10	0.09	0.09
Pooled EER	3.85	3.99	3.51	3.50

EER for 5 attacks (A7, A8, A9, A11 and A16) in comparison with the baseline detection system. This is highlighted in Table 2 using red color for worse EER, blue color for better EER, and black color for nearly the same EER.

The IA/IF-KNN detector with the proposed pre-processing stage gives a pooled EER and min-t-DCF of respectively 3.50% and 0.09% which is roughly the same compared to the baseline system's EER and min-t-DCF. While comparing the performance for each individual attacks (A7-A19), the proposed method improves the EER for 3 attacks (A17, A18, and A19) and gives worse EER for 10 attacks (A7-A16) in comparison with the baseline detection system.

For the IA/IF-KNN system, all VC attacks (A17-A19) have lower EER with the pre-processing stage than without it. For A17, A18, and A19 attacks, EER is reduced to 4.56%, 4.46%, 3.50% from 11.90%, 10.15%, 6.63% respectively.

Of particular interest is attack A17 where we note that "...this method was judged to have the highest spoofing capability in Voice Conversion Challenge 2018" [25]. More recently in [5] (see Table 3) A17 is generally considered the worst attack with the best performing system (Capsule Network) reporting an EER of 3.76% on this attack. For two systems in this work, EER is substantially improved to 3.97%, 4.56% from 7.47%, 11.90% respectively for the

LFCC-GMM, IA/IF-KNN systems. With the exception of the Capsule Network (LFCC+Deep Learning) which reports EER of 3.76% for A17, the systems in this work with the pre-processing stage perform better than all the other systems, i.e. Res-TSSDnet (6.01%), ResNet18-LCML-FM (6.19%), LCNN-LSTM-sum (9.24%), ResNet18-OC-Softmax (9.22%), ResNet18-AM-Softmax (13.45%), ResNet18-GAT-T (28.02%), ResNet18-GAT-s (21.74%) and PC-DARTS (30.20%) where the A17 attack is the worst attack [5]. Performance of these systems for the A17 attack, may be improved with the proposed pre-processing stage.

7 Related work

The various LA spoofing detection methods developed differ by the front-end features used to acquire discriminative information and by the back-end classifiers used for generating the decision score based on which genuine/spoof speech classification is performed. The promising features used for spoofing detection are especially but not limited to Constant-Q Cepstrum Coefficients (CQCC), LFCC, MFCC, Inverse Mel frequency Cepstrum Coefficients (IMFCC), and neural network embedding [19, 26–29]. In some mechanisms, the above-mentioned features are used in combination with source features such as epochs, peak to side lobe ratio to obtain the complementary information that aids detection [18, 30, 31]. The conventional feature extraction is carried out in the frequency domain using filter banks to obtain the short-term sub band spectral features. Some DNN based spoofing detection methods use these features to extract the network embeddings that serve as the feature for categorization [32].

Despite the information richness, the time domain is not considered generally for countermeasure except in a few cases. These include the processing of incoming speech signals in the time domain to separate out the temporal dependency feature which is used in conjunction with the source features for detection [33], temporal convolution for spoofing detection [30] and the usage of variation in temporal distribution of amplitudes for genuine and spoofed speech for classification. The statistical features of IA and IF derived using EMD are calculated along time domain in [24] for spoofing detection. The raw speech waveform is used as input in some DNN based spoofing detectors and this has been made possible by using sinc filters [5, 34, 35].

Rather than extracting features from either time or frequency domain, some spoofing detection methods have used a combined approach [32, 35, 36]. Here the spectral and temporal domains are combined either at feature level by combining the features extracted form both domains or at score level by fusing the individual scores of classifiers by using the features extracted from each domain or by combining the intermediate feature representations of domains within the detector model itself.

In all these spoofing detection approaches, the speech signal is considered as whole without considering the level of impact on various types of segments within a speech signal. That is how our work differs from others where the voiced

and unvoiced segments of a speech signal is separately analyzed to quantify the impact of spoofing on them and used that for spoofing detection.

8 Discussion

Further elaborating for the pooled EER and worst attack performance, we refer the reader to [5]. The IA/IF-KNN system with the pre-processing stage, has better pooled EER performance and better performance against the worst attack than systems ResNet18-GAT-T, ResNet18-GAT-s, PC-DARTS, and RawNet2. With the exception of Capsule Network, the other systems have better pooled EER but worse performance on the worst attack than the worst attack (A17) on IA/IF-KNN system with the pre-processing stage.

For the systems considered in this paper (LFCC+GMM and IA/IF+KNN), in general the pooled results (t-DCF and EER) are roughly the same when using the entire speech signal or with the pre-processing stage (voiced segments and downsampled) where we note that with the pre-processing stage we use approximately 1/4 of the signal samples. We measured execution time for our implmentations of the LFCC+GMM and IA/IF+KNN baseline systems and compared to the systems with the proposed pre-processing stage. We find that, including the overhead for the voiced/unvoiced detector, execution times are reduced by 1.86%, 1.95% for the LFCC+GMM, IA/IF+KNN respectively when using the pre-processing stage.

9 Conclusions

In this paper, we present our observations that the largest spectral differences between bonafide and spoofed speech, lie in the 0-4 kHz band of voiced speech segments. Based on this observation, we propose a pre-processing stage which subsamples voiced frames prior to spoofing detection. The application of the proposed method to the LFCC+GMM and IA/IF+KNN systems reduces the input speech data while maintaining similar pooled EER as that for the baseline systems. Furthermore, our results show substantial improvements in the detection accuracy by both the systems for A17 voice conversion attack and in the A18 and A19 voice conversion attacks for the IA/IF+KNN system. The ASVspoof 2019 A17 voice conversion attack is recognized to have one of the highest spoofing capabilities and has the worst EER for most of the top performing spoofing detectors. We note that the proposed pre-processing stage reduces the speech data by approximately a factor of 4, due to subsampling and using only voiced frames, which may be important for resource-constrained spoofing detectors. Although only two systems were considered, this pre-processing stage may be beneficial to other systems as well.

Acknowledgement

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 19/FFP/6775 and 13/RC/2077_P2.

References

1. Z. Wu and H. Li, "On the Study of Replay and Voice Conversion Attacks to Text-Dependent Speaker Verification," *Multimed. Tools Appl.*, vol. 75, no. 3, pp. 1–17, 2015.
2. J. Lindberg and M. Blomberg, "Vulnerability in Speaker Verification-a Study of Technical Impostor Techniques," in *Proc. Euro. Conf. Speech Commun. Tech. (Eurospeech)*, 1999, pp. 5–9.
3. Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, and M. Sahidullah, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 09 2015.
4. M. Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1008–1012.
5. W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 22–28.
6. T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.
7. J. Lovekin, R. Yantorno, K. Krishnamachari, D. Benincasa, and S. Wenndt, "Developing usable speech criteria for speaker identification technology," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, vol. 1, 2001, pp. 421–424.
8. C. Veaux, J. Yamagishi, and K. MacDonald, *VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.
9. T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *arXiv preprint arXiv:1804.09618*, 2018.
10. A. Consortium, "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf, 2019.
11. L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice Hall, 1978.
12. B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00587
13. X. Wang et al., "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474

14. B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 132–157, jan 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2020.3038524
15. K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639317303710
16. C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-m. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 12 2016, pp. 1–6.
17. W.-C. Huang, Y.-C. Wu, K. Kobayashi, Y.-H. Peng, H.-T. Hwang, P. L. Tobing, Y. Tsao, H.-M. Wang, and T. Toda, "Generalization of spectrum differential based direct waveform modification for voice conversion," 2019. [Online]. Available: https://arxiv.org/abs/1907.11898
18. X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU System for ASVspoof 2015 Challenge," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015.
19. M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 77–81.
20. P. L. De Leon and B. Stewart, "Synthetic Speech Detection Based on Selected Word Discriminators," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, 2013.
21. S. H. Mankad and S. Garg, "On the Performance of Empirical Mode Decomposition-Based Replay Spoofing Detection in Speaker Verification Systems," *Prog. Artif. Intell.*, vol. 9, pp. 325–339, 2020.
22. H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 10 2020, p. 1844.
23. P. Tapkir and H. Patil, "Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 721–725.
24. A. S. M S, P. L. De Leon, S. Sandoval, and U. Roedig, "Low-complexity speech spoofing detection using instantaneous spectral features," in *29th International Conference on Systems, Signals and Image Processing (IWSSIP 2022)*, 06 2022. [Online]. Available: https://hdl.handle.net/10468/13215
25. T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z.-H. Ling, "A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment," in *Proc. Odyssey 2018*, 06 2018, pp. 187–194.
26. H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4633–4644, 2018.
27. M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: An

evaluation on asvspoof 2015," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2016.
28. G. Lavrentyeva et al., "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1033–1037.
29. B. Chetttri et al., "Ensemble Models for Spoofing Detection in Automatic Speaker Verification," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1018–1022.
30. X. Tian, X. Xiao, E. S. Chng, and H. Li, "Spoofing Speech Detection Using Temporal Convolutional Neural Network," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.
31. S. Jelil, R. K. Das, S. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH*, 2017.
32. H. Tak, J. weon Jung, J. Patino, M. Todisco, and N. W. D. Evans, "Graph attention networks for anti-spoofing," in *Interspeech*, 2021.
33. M. Witkowski, S. Kacprzak, P. Żelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 27–31.
34. H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, 06 2021, pp. 6369–6373.
35. H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 09 2021, pp. 1–8.
36. J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, 05 2022, pp. 6367–6371.