Low-Complexity Speech Spoofing Detection using Instantaneous Spectral Features

Arun Sankar M. S.
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
asankar@ucc.ie

Phillip L. De Leon
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, New Mexico, U.S.A.
pdeleon@nmsu.edu

Steven Sandoval
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, New Mexico, U.S.A.
spsandov@nmsu.edu

Utz Roedig
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
u.roedig@ucc.ie

***Abstract*—Over the last decade, various detection mechanisms for spoofed speech have been proposed. Thus far the development focus has been on detection accuracy, largely ignoring secondary goals such as computational complexity or storage effort. In this work, we use empirical mode decomposition to compute intrinsic mode functions which are then demodulated to obtain features consisting of short-time statistics of instantaneous amplitude and instantaneous frequency. These features are then used with a simple k-nearest neighbours classifier. We further show that voiced segments from short speech signals can be used in the feature extraction resulting in a spoofing detection competitive with top-performing systems while having up to $103\times$ less computation.**

***Keywords*—Computer security, Biometrics, Speaker recognition, Speech processing**

I. Introduction

Automatic Speaker Verification (ASV) systems are popular as a low-cost and flexible technology for biometric authentication. 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].Countermeasures to detect spoofed speech and thus prevent an attack, are in active development and the ASVspoof biannual challenge, initiated in 2015, has assisted with advancing the research through organized trials and evaluations [2]. The fourth challenge organized recently in 2021 focused on discriminating between genuine and spoofed or deepfake speech. At the time of writing this paper, ASVspoof 2021 challenge results were not yet available, hence we present and compare our work with ASVspoof 2019 challenge results [3]. The development focus has thus far been on detection accuracy, largely ignoring secondary goals such as computational and storage effort. However, voice control is now added to many digital systems such as smart speakers (Amazon Alexa), mobile phones (Siri) and cars (Jaguar). As ASV and spoofing detection are to be used at scale, either implemented on many small embedded devices or in a cloud back-end infrastructure, complexity must be minimised to save resources.

In this paper we propose a novel Empirical Mode Decomposition (EMD) based spoofing detection system which has low computational complexity and storage requirements while providing a high detection accuracy. We use short-time statistics of Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) from the Intrinsic Mode Functions (IMFs) resulting from the EMD as features which are fed to a simple k-Nearest Neighbours (KNN) classifier. We then demonstrate that classification accuracy can be improved when only using the voiced regions of the speech signal. Finally, as the complexity of the EMD is dependant on signal length we show that a short duration of 2s is sufficient. Our evaluation shows that the proposed method can compete with state-of-the-art spoofing detection mechanisms as described in ASVspoof 2019 challenge [3] while reducing computational complexity by a factor of up to $103\times$ and having negligible storage requirements.

The contributions of this work are: (i) a low complexity EMD based spoofing detector; (ii) a novel EMD based feature vector for spoofing detection; (iii) a demonstration that the voiced components of a speech signal are best for spoofing detection; and (iv) an analysis of the impact of signal length on spoofing detection accuracy.

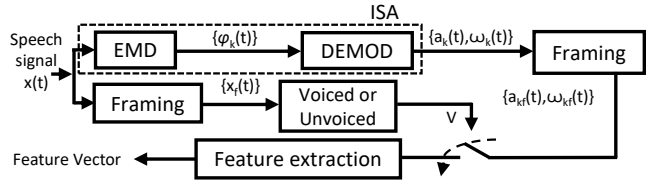
II. Empirical Mode Decomposition and Proposed Features

Huang's original definition of the Hilbert spectrum uses EMD to determine a set of IMFs which are individually demodulated with the Hilbert Transform (HT) to obtain IAs

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 19/FFP/6775. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

978-1-6654-9578-3/22/$31.00 ©2022 IEEE

Fig. 1. Block diagram of the EMD based feature vector creation for spoofing detection.



$\{a_k(t)\}$ and IFs $\{\omega_k(t)\}$ [4]. This analysis is also known as the Hilbert-Huang Transform (HHT) [4]. EMD has been used before in the context of spoofing detection. However, existing work has used EMD only as replacement for Mel frequency Cepstrum Coefficients (MFCC) filter banks [5] and at a pre-processing stage to reconstruct the signal using a subset of IMFs [6]. In this work we directly construct feature vectors from the IMF removing traditional steps such as filter banks and MFCC.

A. Empirical Mode Decomposition

The EMD algorithm sequentially decomposes the input signal into a set of IMFs $\{\varphi_k(t)\}$ by iteratively calling the sifting algorithm [4]. Several improvements have been proposed to improve the sifting algorithm [7], [8]. However, IMFs are not orthogonal to each other and as a result, a decomposition into IMFs is not unique. Due to this ambiguity, the decomposition returned by EMD does not always capture the assumed/true underlying signal components as expected. More specifically, undesirable effects termed mode mixing [9] and component splitting [10] may be present in the decomposition.

In addition to the improvements proposed to the sifting algorithm, several researchers have also proposed variations on the original EMD algorithm. For example, the Ensemble Empirical Mode Decomposition (EEMD) [9] introduced ensemble averaging in order to address the mode mixing problem via an additive noise and an averaging of IMF estimates. In our prior work [11], we have also proposed improvements to Complementary Ensemble Empirical Mode Decomposition (CEEMD) including: 1) a modification to the ensemble averaging which guarantees that the average IMF is a true IMF [9] and 2) a change from the additive noise used in ensemble averaging to a complimentary pair of narrowband tones [12] which we termed "tone masking". For a clear presentation of EMD and the sifting algorithm, the reader is referred to [11].

B. Feature Extraction

In this work, we propose to use short-time statistics of the IA/IF from the IMF resulting from EMD. The HT approach to demodulation is used in the HHT, however, others have proposed alternatives to the HT in order to improve local behavior [7]. In [11], the authors proposed numerical stabilization techniques for Huang's iterative IA estimation and direct IF estimation algorithms which give good estimates for the IA/IF parameters. MATLAB® codes for EMD and IA/IF estimation of resulting IMFs may be found at [13], [14].

Fig. 1 shows a block diagram of the proposed feature extraction process. We begin by estimating the IA/IF parameters $\{a_k(t), \omega_k(t)\}$ from the first ten IMFs that result from the sifting operation in EMD of the speech signal $x(t)$. It is observed that the speech signals under investigation, in general, yielded ten IMFs. Next, feature vectors are formed as follows. First, $\{a_k(t), \omega_k(t)\}$, $1 \leq k \leq 10$ are framed into 20ms windows and the mean, variance, skewness, and kurtosis are computed resulting in eight statistics per IMF per frame. Second, for each IMF, the statistics are then averaged over the frames. Finally, the eight statistics from each IMF are stacked resulting in a feature vector of length 80. In the case where fewer than ten IMFs result from EMD, we zero pad the feature vector.

Phonemes in a speech signal are created by the vocal cords and the vocal tract. Voiced speech is created when the vocal cords vibrate while a phoneme is pronounced. Unvoiced speech does not make use of vocal cords. We label the frames as voiced or unvoiced using Zero Crossing Rate (ZCR). The use of statistics only stemming from voiced speech when averaging statistics per IMF improves detection results as we will show in our evaluation.

III. Experiments and Evaluation

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. The training and development subsets of LA are used for conducting experiments related to the development of the detection model while the evaluation set is utilized for measuring detection performance. For additional information, please see [2].

The evaluations includes: 1) examination of spoofing detection accuracy using voiced or unvoiced speech segments; 2) examination of spoofing detection accuracy as function of speech signal length; 3) comparison of the detection accuracy with state-of-the-art algorithms as described in ASVspoof 2019 challenge; 4) analysis of computation complexity and storage requirements.

A. Results with Voiced vs. Unvoiced Speech

For the first experiment we consider feature extraction from voiced or unvoiced speech segments that is, the average IA/IF statistics are computed only for voiced or unvoiced speech segments. Our motivation in considering voicing is based on fundamental differences in the production and properties of voiced and unvoiced segments. We consider the simple classifiers KNN, Neural Network (NN), and Support Vector Machines (SVM). The NN Classifier has a single hidden layer with 10 neurons and SVM does classification by selecting the most matching hyper plane using quadratic kernel. We partition the ASVspoof 2019 training set data for each spoofing algorithm, i.e. A01 - A06 into 80% for classifier training and 20% for testing. We note that classifiers are trained for each spoofing algorithm. The average results of 100 trials of randomized partitioning are determined.

TABLE I

SPOOFING DETECTION RESULTS FOR THE PROPOSED FEATURE VECTOR EXTRACTED FROM THE FULL SPEECH SIGNAL AS WELL AS VOICED AND UNVOICED SEGMENTS FROM THE ASVSPOOF 2019 ALGORITHMS A01-A06. ON AVERAGE, KNN CLASSIFICATION USING FEATURE VECTORS EXTRACTED FROM VOICED SPEECH, PERFORMED BEST.

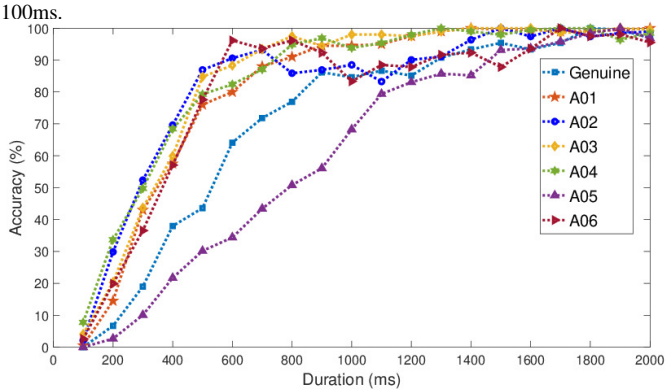
Spoofing	**Full(%)**			**Unvoiced (%)**			**Voiced (%)**		
Algorithm	**KNN**	**NN**	**SVM**	**KNN**	**NN**	**SVM**	**KNN**	**NN**	**SVM**
A01	95.9	96.8	96.4	89.0	88.2	90.9	97.2	96.5	**97.6**
A02	98.5	**98.7**	97.6	77.8	79.6	82.9	96.0	**98.7**	**98.7**
A03	96.7	97.7	97.6	81.5	84.2	87.0	97.2	98.1	**98.3**
A04	92.0	92.7	93.9	60.2	59.1	61.2	93.5	93.3	**95**
A05	93.2	96.7	96.0	72.8	76.0	78.3	95.2	97.2	**97.6**
A06	76.4	79.1	77.9	66.7	68.1	70.9	**95.0**	78.4	82.1
Average	92.1	93.6	93.2	74.7	75.9	78.5	**95.7**	93.7	94.8

As a baseline, detection accuracy using the proposed feature vector extracted from the full speech signal (voiced and unvoiced segments), is provided in Table I columns 2-4 and along with results using only the voiced segments (columns 5-7) or only unvoiced segments (8-10). We find that in general: 1) classifiers using feature vectors extracted from voiced speech segments perform better than when using unvoiced speech segments, 2) classifier accuracy is lower for A04-A06 (waveform concatenation and voice conversion spoofing algorithms) than with A01-A03 (Text-to-Speech (TTS) spoofing algorithms), and 3) With the exception of A06, all three classifiers have similar accuracy, however, KNN performs best with A06. Although we have not investigated further, other proposed features extracted from voiced speech may improve accuracy with other detection algorithms. In the subsequent work presented in this paper, we use feature vectors extracted only from voiced speech segments and the KNN classifier since, on average, this performed best.

B. Detection Accuracy as a Function of Signal Length

The duration of the speech signal under analysis affects the quality of the feature vector and has also a direct impact on the computational complexity (We discuss computation complexity separately in Section III-D). To better understand the role of signal length, experiments are conducted using three sets of speech signals with an approximate duration of 2s, 4s, and 6s from the training data set of each of the spoofing algorithms (A01-A06). The EMD analysis is performed on each of the three sets of speech signals and features are extracted over lengths which are increasing by 100ms starting with a length of 500ms. Detection accuracy is normalized against the KNN voiced results in Table I, meaning that 100% implies results are equal to when using the entire speech signal. Results of detection accuracy as a function of signal length are shown in Fig. 2 for speech duration of 2s; although not shown results using 4s and 6s speech signals are similar to 2s. We find that the relative accuracy increases with speech signal duration and reaches saturation around 2s for all the data sets of spoofing algorithms. Thus with the proposed feature vector extracted from the voiced segments of at least a 2s signal, accurate detection of spoofed speech is possible. Other speech processing applications such as speaker recognition [15] have also considered speech signal length. However, to the best of our knowledge spoofing detection with short signals has not been previously explored but is important for resource-limited scenarios or cases where spoofing detection with short delay is required.

Fig. 2. Relative detection accuracy as a function of signal length for genuine and spoofed speech. In this plot, EMD is performed over speech signals of length 2s and features are extracted over lengths which are increasing by 100ms.



C. ASVspoof 2019 Evaluation and Results

For the third experiment, we developed models using both the training and development sets of ASVspoof 2019 challenge. Following the ASVspoof 2019 challenge evaluation plan [2], the model is tested on the evaluation data set to analyze detection performance on unknown attacks in addition to the known attacks (A16 and A19). The results of the experiment are given in Table II using t-DCF and EER metrics as defined by ASVspoof. Our model, using the proposed feature vector extracted from voiced segments using a KNN classifier ranks in the top five (see Table III). We note that systems T05, T45, T60, T24, and T50 are all ensemble classifiers unlike the proposed system which uses a single, simple KNN classifier. The proposed model performs well on all the spoofing attacks except A17, A18, and A19. Among these, A19 is a known attack that uses the same algorithm as A06 but with different data hence the similar performance to A06. The same reasoning explains the performance of the proposed model on known attack A16 that uses the same algorithm as A04.

D. Computation Complexity and Storage Requirements

Among the best performing systems in the ASVspoof 2019 challenge, only the architecture details for T45 [16] and T60 [17] systems are available. T45 is a fusion of four Deep Neural Networkss (DNNs) that each differ in the front-end features. By our estimate, the detector requires a total of 40.8M parameters and at least 3.99G Multiply And Accumulate (MAC) units where we have only counted MAC units for the convolution layers that contribute significantly to the computation. The T60 is an ensemble model consisting of four DNNs, an i-vector based SVM, and two Gaussian Mixture Model (GMM) classifiers that use MFCCs and inverse MFCCs as features. By our estimate, the DNNs require 2.59M parameters and 604.22M MAC units.

TABLE II
PROPOSED COUNTERMEASURE'S EQUAL ERROR RATE (EER) AND TANDEM-DECISION COST FUNCTION (T-DCF) RESULTS FOR SPOOFING ALGORITHMS IN ASVSPOOF2019.

	A07	**A08**	**A09**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**	**Pooled**
EER	2.09	2.09	2.09	2.09	2.09	2.09	2.09	2.09	2.09	2.09	11.90	10.15	6.63	3.51
t-DCF	0.0509	0.0542	0.1732	0.0528	0.0563	0.0516	0.0502	0.0515	0.0563	0.0594	0.7489	0.3370	0.1564	0.0953

TABLE III
RESULTS OF THE PROPOSED COUNTERMEASURE WITH THE TOP PERFORMING AND BASELINE SYSTEMS FROM THE ASVSPOOF 2019 CHALLENGE. IN TERMS OF POOLED EER AND T-DCF VALUES, THE PROPOSED SYSTEM RANKS IN THE TOP FIVE.

System	**EER (%)**	**t-DCF (%)**
T05	0.22	0.0069
T45	1.86	0.0510
T60	2.64	0.0755
T24	3.45	0.0953
Proposed System	**3.51**	**0.0953**
T50	3.56	0.1118
B01: LFCC - Baseline system	8.09	0.2116
B02: CQCC - Baseline system	9.57	0.2366

As depicted in Fig. 1, the computational complexity of the proposed method is dominated by EMD; we consider spoofing detection of 2s signals and an average number of sifting iterations per IMF for a total of ten IMFs. Following [18], [19], we estimate the EMD requires 38.40M MAC units and negligible requirement for parameter storage. Taken together, the proposed system requires far fewer computational resources and storage, as compared to T45 and T60, which makes it a suitable counter measure for low resource scenarios such as client-side detection in personal voice assistants.

IV. CONCLUSIONS

In this paper we have proposed a new countermeasure which uses features consisting of short-time statistics of IA/IF parameters of IMFs resulting from the EMD and a simple KNN classifier. The proposed spoofing countermeasure ranks 5^{th} in terms of detection performance (EER and t-DCF) compared with algorithms reported in ASVspoof 2019. However, all better performing algorithms use ensemble classifiers which are generally demand significantly more computational effort and storage space. We are certainly aware that DNN based models can learn features and can outperform our proposal in terms of detection performance. However our aim was to reduce the computation complexity and our method requires $103\times$ and $15\times$ less computation than T45 and T60 algorithms respectively. The proposed resource-efficient method may be used as an alternative or to augment published ensemble classifiers. Given the rise of personal voice assistants such as Alexa, Siri and Cortana, spoofing detection is now required at scale. A better balance of resource efficiency and detection accuracy is required. Finally, we investigated feature extraction from only voiced segments and found this improves detection accuracy and we found accurate detection was possible with signals as short as 2s. Spoofing detection with short signals has not been previously explored but is important for resource-limited scenarios or if spoofing detection with short delay is required.

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] ASVspoof Consortium, "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf, 2019.

[3] 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.

[4] N. E. Huang et al, "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis," *Proc. R. Soc. London Ser. A*, vol. 454, no. 1971, pp. 903–995, 1998.

[5] 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.

[6] 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.

[7] R. Rato, M. D. Ortigueira, and A. Batista, "On the HHT, Its Problems and Some Solutions," *Mech. Syst. and Sig. Process.*, vol. 22, no. 6, pp. 1374–1394, 2008.

[8] S. Sandoval, M. Bredin, and P. L. De Leon, "Using Linear Prediction to Mitigate End Effects in Empirical Mode Decomposition," in *Proc. IEEE Global Conf. Sig. Info. Process. (GlobalSIP)*, 2018, pp. 281–285.

[9] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method," *Adv. Adapt. Data Anal.*, vol. 1, no. 01, pp. 1–41, 2009.

[10] S. Sandoval, M. Bredin, and P. L. De Leon, "Dominant Component Tracking for Empirical Mode Decomposition using a Hidden Markov Model," in *Proc. IEEE Global Conf. Sig. Info. Process. (GlobalSIP)*, 2018, pp. 116–120.

[11] S. Sandoval and P. L. De Leon, "Advances in Empirical Mode Decomposition for Computing Instantaneous Amplitudes and Instantaneous Frequencies," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, 2017, pp. 4311–4315.

[12] N. Senroy, S. Suryanarayanan, and P. F. Ribeiro, "An Improved Hilbert–Huang Method for Analysis of Time-Varying Waveforms in Power Quality," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1843–1850, 2007.

[13] "MATLAB Signal Processing Toolbox," 2021, the MathWorks, Natick, MA, USA.

[14] S. Sandoval, "Instaneous Spectral Analysis," https://github.com/ssandova/ISA-public, 2018.

[15] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2011, pp. 2341–2344.

[16] G. Lavrentyeva et al., "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1033–1037.

[17] 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.

[18] Y.-H. Wang, C.-H. Yeh, H.-W. V. Young, K. Hu, and M.-T. Lo, "On the computational complexity of the empirical mode decomposition algorithm," *Physica A, Stat. Mech. its Appl.*, vol. 400, pp. 159–167, 2014.

[19] S. Sandoval and P. L. De Leon, "Advances in empirical mode decomposition for computing instantaneous amplitudes and instantaneous frequencies," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, 2017, pp. 4311–4315.