Energy Separation Algorithm based Features for Replay Spoof Detection

Madhu R. Kamble

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India

madhu_kamble@daiict.ac.in

Abstract

Voice biometric systems are vulnerable to various kinds of spoofing attacks. Replay spoofing attack is one of the more easiest way to spoof the biometric system. Replay attack requires the pre-recorded speech samples of the genuine speaker without having prior knowledge of sophisticated speech algorithms. In this paper, we develop a countermeasure using Energy Separation Algorithm (ESA) to detect the difference between natural and replayed speech signal. The ESA-based proposed Instantaneous Amplitude and Instantaneous Frequency Cepstral Coefficients (i.e., IACC and IFCC) feature set gave lower Equal Error Rate (EER) compared to the given baseline system of Constant Q Cepstral Coefficients (CQCC). On evaluation set of ASV Spoof 2017 Challenge Version 2.0 database, the EER obtained with IACC and IFCC are 12.00 % and 12.79 %, respectively. To explore the complementary information of IACC and IFCC, we performed score-level fusion and reduced the EER to 9.64 % on eval dataset.

Index Terms: Speaker Verification, Spoofing, Replay, Teager Energy Operator, Energy Separation Algorithm.

1. Motivation

Now-a-days the use of biometric patterns, such as face, iris, fingerprint, etc. are widely used in many civilian and private-sector applications for personal recognition [1]. To overcome from critical password protection (for example: PINS, patterns, passwords, etc.), biometric passwords are easy to protect our applications [2]. Biometrics cannot be lost or forgotten, they are difficult for attackers to forge [3]. Voice biometric can be considered either as an anatomical or as a behavioral characteristic. The goal of the Automatic Speaker Verification (ASV) system is to determine or verify the identity of an individual speaker's voice [4]. Among current concerns of a threat to the systems, one of the vulnerabilities is *spoofing* and it is defined as, the speaker who masquerade as the target speaker to gain the access to protected data [5, 6]. In the literature, there are various kinds of spoofing attacks, namely, speech synthesis (SS) [7], voice conversion (VC) [8], replay [9], twins and impersonation [10]. A general Spoof speech detection system is shown in Figure 1.

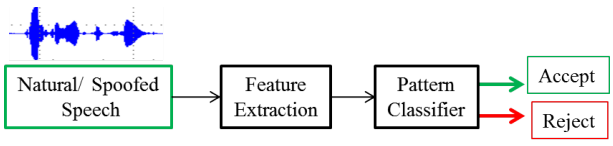

Figure 1: *Spoofing detection framework.*

Various approaches were proposed for detection of spoofing algorithm on various datasets and hence, there is a need to provide a common platform and have a common performance metric to evaluate the spoofing countermeasures [6]. The first spoofing and countermeasures challenge focusing on SS and VC spoofing attacks was organized as a special session in INTERSPEECH 2015 [11], where as the second challenge was focused on replay attacks and was organized in INTERSPEECH 2017 [12]. Researchers proposed several countermeasures at the feature and classifier side on both ASV Spoof 2015 and 2017 Challenge database. For spoof speech detection (SSD) task features such as, Constant Q Cepstral Coefficients (CQCC) [13, 14], Linear Frequency Cepstral Coefficients (LFCC) [15], Mel Frequency Cepstral Coefficients (MFCC) [16], [17] are state-of-the-art features with simple Gaussian Mixture Model (GMM) classifier. In this doctoral work, we are developing a countermeasure and proposed Energy Separation Algorithm-based feature sets, namely, IACC and IFCC and evaluated on ASV Spoof 2017 challenge database for replay SSD task.

2. Key Research Issues

In the area of replay detection many of the issues were observed and they were addressed as follows. As replay speech signals are the signals obtained with the convolution of original speech signal and the impulse response of the intermediate device, environment [18]. Because of the insertion of intermediate device characteristics the spectral energies are emphasized in the higher frequencies regions as shown in Figure 2 and it is also reported in [19, 20]. The study with the effect of Cepstral Mean Variance Normalization (CMVN) is also reported in [21–23]. Various acoustics features were also reported in [19, 20, 24–27]

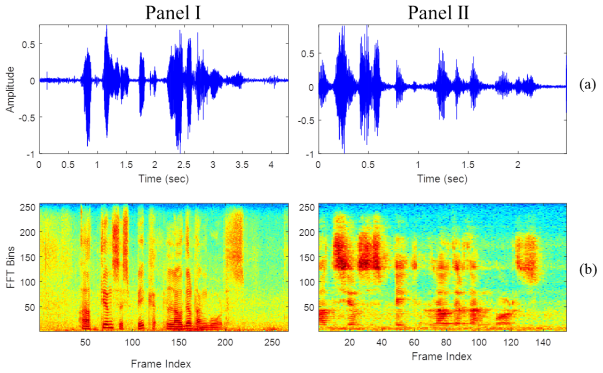

Figure 2: *Spectral energy densities of Panel I: Natural and Panel II: Replayed speech signal. (a) time-domain speech signal, and (b) spectral energies, respectively.*

3. Major Contributions

The major contribution towards this research problem is addressed by proposing various acoustic cepstral features using the Teager Energy Operator (TEO) [28, 29]. The study reported

in [26, 30–32] uses the demodulation of TEO profile. The energy of TEO $\Psi_d\{x(n)\}$ is separated into its Instantaneous Amplitude (IA ($a_i[n]$)) envelope and Instantaneous Frequency (IF ($\Omega_i[n]$)) [33], [34] and they are given by:

$$\Psi_d\{x(n)\} = x^2(n) - x(n-1)x(n+1) \approx a_i^2[n]\Omega_i^2[n].$$

$$a_i[n] \approx \frac{2\Psi_d\{x[n]\}}{\sqrt{\Psi_d\{x[n+1]) - x[n-1]\}}},$$

$$\Omega_i[n] \approx arcsin\sqrt{\frac{\Psi_d\{x[n+1] - x[n-1]\}}{4\Psi_d\{x[n]\}}}.$$

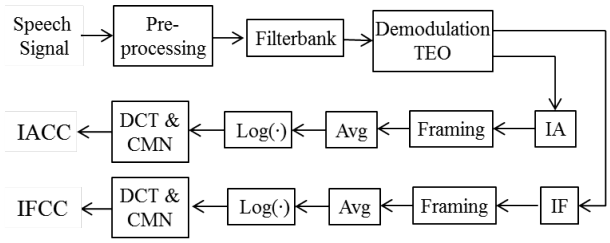

Figure 3: *Schematic block diagram of proposed demodulation-based feature sets.*

The schematic diagram of the proposed feature extraction is shown in Figure 3. The input speech signal is passed through a pre-emphasis filter to strengthen the higher frequencies and this emphasized signal is given to the filterbank. Here, we have used Gabor filterbank to obtain number of subband filter speech signal as TEO works on monocomponent signals. These number of subband filtered signals are then passed to TEO to extract energy traces of those subband filtered signals. These energy traces are further separated using Energy Separation Algorithm (ESA) and gives the Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) of respected subband filtered signals. Now, the extracted IA and IF are processed into frames with 20 ms of window length and having a window shift of 10 ms. After obtaining the frames of each signal they are averaged and are followed by logarithm operation. Furthermore, Discrete Cosine Transform and CMN method are used to retain first few Instantaneous Amplitude Cepstral Coefficients (IACC) and Instantaneous Frequency Cepstral Coefficients (IFCC).

4. Discussion

The experiments are done on the ASV Spoof 2017 Challenge version database [12], [35]. The Challenge organizers provided a baseline system that contains CQCC feature set and Gaussian Mixture Model (GMM) as classifier with 512 number of Gaussian components in GMM. The baseline system gave an Equal Error Rate (EER) on development and evaluation set of 10.35 % and 28.48 %, respectively. However, the baseline system do not perform better for replay detection task and hence, cannot be used as a good countermeasures. The proposed IACC and IFCC feature set results are shown in Table 1. The proposed feature sets are extracted using 40 subband filtered signals obtained from Gabor filterbank having filter bandwidth of 200 Hz. The feature vector includes 40 static coefficients appended along with their delta and double-delta feature vector resulting in 120-dimensional feature coefficients. The results obtained from IACC and IFCC feature set have lower EER compared to the baseline system resulting in 6.48 % and 4.12 % on dev set whereas on eval set the EER obtained are 12.00 % and 12.79 %, respectively. To explore the possible complementary information present in both the proposed feature set, we used score-level fusion of those feature set and obtained a reduced EER to 2.01 % and 9.64 % on dev and eval dataset, respectively. The performance is also shown by the DET curves in Figure 4 (a) for dev set and Figure 4 (b) for eval set with CQCC, IACC,IFCC feature sets along with score-level fusion of IACC and IFCC.

Table 1: *Comparison of best proposed feature set with other feature set on development (Dev) and evaluation (Eval) dataset*

Feature Sets	Dev	Eval
CQCC (Baseline system)	10.35	28.48
IACC	06.48	12.00
IFCC	04.12	12.79
IACC+IFCC	**02.01**	**09.64**

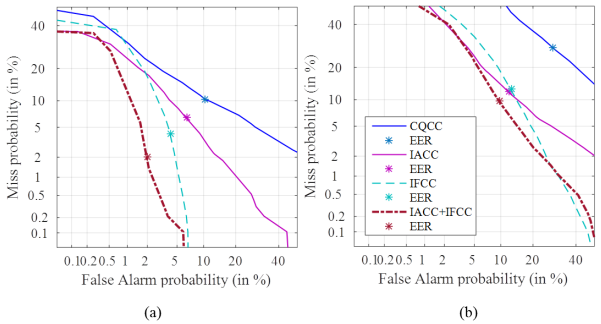

Figure 4: *Individual DET curves of CQCC, AM-FM features. (a) on dev and (b) eval set.*

5. Future plans, Summary and Conclusions

Though several countermeasures are approached towards replay detection task, however they are restricted to particular database. If the test set is changed or different condition are introduced in test data the performance might be changed. To solve this kind of issue a strong countermeasure is required to develop. However, the spoofing attacks consist of 4 different attacks and hence, there is a need to develop a generalized countermeasure in this research area. Given the above discussed issues in spoofing, our future plan will be to develop a generalized countermeasure with a joint protocol of spoofing and speaker verification.

In this doctoral work, we studied the demodulation-based features to detect natural *vs.* replayed spoofed speech. The computation of Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) from Teager Energy Operator Energy Separation Algorithm (TEO-ESA) was affected by the parameters of filter, namely, shape of filter, choice of bandwidth, time resolution, etc.

6. Acknowledgements

Author would like to thank the organizing committee members of INTERSPEECH 2018 for providing the platform to present our doctoral work. In addition, author also thank her supervisor Prof. Hemant A. Patil for his continuous support and motivation throughout the research and finally thank University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF) and authorities of DA-IICT Gandhinagar.

7. References

[1] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[2] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[3] B. Beranek, "Voice biometrics: success stories, success factors and what's next," *Biometric Technology Today*, vol. 2013, no. 7, pp. 9–11, 2013.

[4] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice anti-spoofing," *Handbook of biometric anti-spoofing, S. Marcel, SZ Li, and M. Nixon, Eds. Springer*, 2014.

[5] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[6] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *INTERSPEECH*, Lyon, France, 2013, pp. 925–929.

[7] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[8] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Taipei, Taiwan, China: IEEE, 2009, pp. 3585–3588.

[9] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH, Lyon, France*, 2013, pp. 940–944.

[10] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 2004, pp. 145–148.

[11] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2–6.

[13] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[14] Todisco, Massimiliano and Delgado, Héctor and Evans, Nicholas, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

[15] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

[16] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[17] Z. Wu, J. Yamagishi, T. Kinnunen *et al.*, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[18] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in ASV systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017, pp. 51–55.

[19] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 97–101.

[20] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 27–31.

[21] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 82–86.

[22] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 17–21.

[23] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 102–106.

[24] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 7–11.

[25] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[26] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[27] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *INTERSPEECH 2017*, Stockholm, Sweden, 2017, pp. 22–26.

[28] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Toronto, Ontario, Canada, 1991, pp. 421–424.

[29] M. R. Kamble and H. A. Patil, "Novel energy separation based instantaneous frequency features for spoof speech detection," in *European Signal Processing Conference (EUSIPCO)*, Kos Island, Greece, 2017, pp. 116–120.

[30] M. R. Kamble and H. A. Patil, "Effectiveness of mel scale-based ESA-IFCC features for classification of natural *vs.* spoofed speech," in *B.U. Shankar et. al. (Eds.) PReMI, Lecture Notes in Computer Sciance (LNCS)*. Springer, 2017, pp. 308–316.

[31] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay spoof speech detection," *Accepted in INTERSPEECH*, Hyderabad, India, 2018.

[32] M. R. Kamble, H. Tak and H. A. Patil, "Amplitude and frequency modulation-based features for detection of replayed spoof speech," *Submitted in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[33] P. Maragos, J.F. Kaiser and T.H. Quatieri , "On separating amplitude from frequency modulations using energy operators," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, San Francisco, California, USA, 1992, pp. 1–4.

[34] Maragos, Petros and Kaiser, James F and Quatieri, Thomas F, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[35] H. Delgado *et al.*, "ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements," in *The Speaker and Language Recognition Workshop ODYSSEY*, Les Sables d'Olonne, France, 2018. [Online]. Available: http://www.eurecom.fr/publication/5504, Last Access 28-May-2018