INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Independent Modelling of High and Low Energy Speech Frames for Spoofing Detection

Gajan Suthokumar[1,2], Kaavya Sriskandaraja[1,2], Vidhyasaharan Sethu[1], Chamith Wijenayake[1], Eliathamby Ambikairajah[1,2]
[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]DATA61, CSIRO, Sydney, Australia
g.suthokumar@unsw.edu.au, k.sriskandaraja@unsw.edu.au, v.sethu@unsw.edu.au, c.wijenayake@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

Spoofing detection systems for automatic speaker verification have moved from only modelling voiced frames to modelling all speech frames. Unvoiced speech has been shown to carry information about spoofing attacks and anti-spoofing systems may further benefit by treating voiced and unvoiced speech differently. In this paper, we separate speech into low and high energy frames and independently model the distributions of both to form two spoofing detection systems that are then fused at the score level. Experiments conducted on the ASVspoof 2015, BTAS 2016 and Spoofing and Anti-Spoofing (SAS) corpora demonstrate that the proposed approach of fusing two independent high and low energy spoofing detection systems consistently outperforms the standard approach that does not distinguish between high and low energy frames.

Index Terms: automatic speaker verification, spoofing detection, binary classification, score fusion

1. Introduction

Automatic Speaker Verification (ASV) is used as a voice biometric to verify the claimed speaker from a speech utterance. ASV systems are more vulnerable to spoofing compared to other biometrics due to a lack of face-to-face contact as in e-commerce [1] and mobile banking [2]. In the context of ASV, spoofing is used to manipulate the speaker verification process by recreating the original speaker's voice. Spoofing methods are categorized as speech synthesis (SS), voice conversion (VC), impersonation or replay [3]. Speech synthesis and voice conversion attacks pose high threat due to widely available state-of-the-art open source toolkits, making them both effective and accessible. Techniques to produce synthesized speech have shown rapid advancement in recent years. As an example, the recently introduced Wavenet [4] is able to produce high quality synthesized speech that is close to natural sounding.

A diverse range of spectral features, phase based features, prosodic long term features and combinations of spectral and phase features have been reported to detect speech synthesis and voice conversion attacks [5]. Linear Frequency Cepstral Coefficients (LFCC) [5], Inverse Mel Frequency Cepstral Coefficients (IMFCC) [5], Constant Q cepstral (CQCC) features [6] and scattering cepstral coefficients (SCC) [7] have been shown to be effective front-ends for spoofing detection. Similarly, a diverse range of classifiers have been evaluated for spoofing detection [8] and Gaussian mixture models (GMM) stand out as the most promising approach thus far [8].

Anti-spoofing systems can either be integrated into a speaker verification system or implemented as a standalone system in parallel to an ASV system [3]. Since speaker verification systems and spoofing detection systems may have conflicting optimization criteria, the stand-alone approach has been preferred.

Early standalone spoofed detection systems only modelled voiced regions. Recently, both voiced and unvoiced regions have been employed [5, 9] since discriminative information is present in both. In particular, speech synthesis and voice conversion methods tend to model voiced speech much more accurately than unvoiced speech [10], while artificial speech based on STRAIGHT [11] and Harmonic Noise Model (HNM) [12] vocoders can be detected based on the power spectrum of non-speech regions. In general, speech synthesis and voice conversion algorithms have distinct voiced and unvoiced speech models [13, 14] and this distinction may be exploited for spoofing detection. It has also been shown that STRAIGHT, HNM and sinusoidal modelling methods fail to model unvoiced plosives and stops, introducing artefacts such as the pre-echo effect [15, 16] which may also be exploited for spoofing detection. More recently, unvoiced speech has been modelled with neural networks to reduce artefacts [17]. Broadly, the voiced-unvoiced boundaries are difficult to model and continuous vocoders that aim to reduce these boundary effects are still under active research [18].

In this paper, we build on the following two observations: a) artefacts in unvoiced regions may help detect spoofing attacks; and b) state-of-the-art speech synthesis and voice conversion methods do not treat voiced and unvoiced regions in the same manner. Specifically, we separate high and low energy frames and model the differences between spoofed and genuine speech in these two (high and low energy) regions independently.

2. Proposed Method

Figure 1 shows an overview of the proposed approach where speech frames are initially identified as either high energy (HE) or low energy (LE) frames using a VAD. Here we expect the LE frames to contain unvoiced speech, which is found at the start and end of words (i.e. V-UV and UV-V boundary frames), gaps and silence. GMM based spoofing detection systems are then implemented for both HE and LE frames in parallel. Finally, the log-likelihood ratios from both HE and LE systems are linearly fused to obtain the final decision.

In the experiments reported in this paper, we utilise front-ends based on LFCCs and IMFCCs as they have shown to be effective for spoofing detection [5]. The LFCC features are computed using a linearly spaced filterbank while the IMFCC

http://dx.doi.org/10.21437/Interspeech.2017-836

front-end employs an inverse mel filterbank. Both of these front-ends use a greater number of filters in the high frequency regions compared to other features types such as SCC.

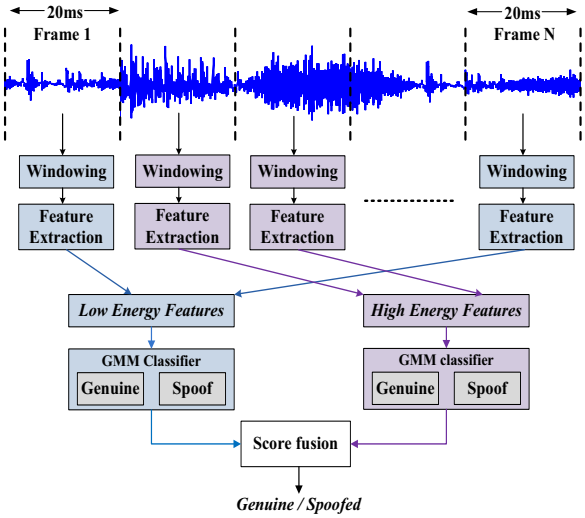

Figure 1: *Overview of proposed independent modelling of low and high energy frames.*

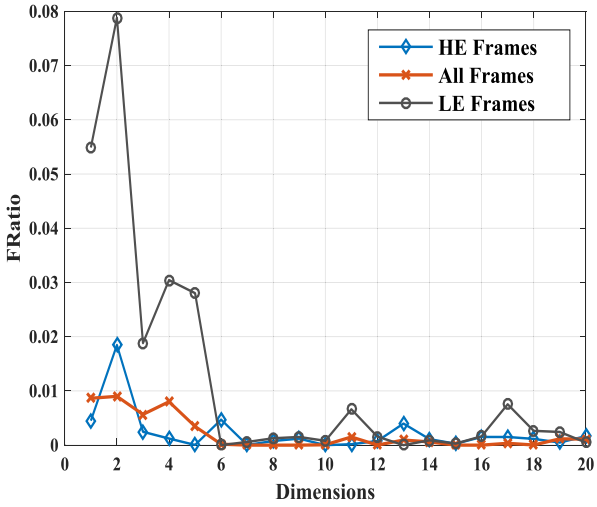

Figure 2: *FRatio computed from IMFCCs for HE frames, LE frames and all frames from the ASVSpoof 2015 training set.*

In order to determine if there are any potential advantages to separating high and low energy frames, we plot Fisher's ratio (FRatio) between IMFCCs from genuine and spoofed speech for both HE and LE frames in Figure 2 and compare them to the FRatio obtained when the frames are not segregated into HE and LE frames (referred to as 'All frames' in Figure 2). It can be seen from the figure that the FRatios obtained using LE frames in particular are much greater than those from HE frames and 'all frames' cases. This suggests that separating the frames into high and low energy ones and modelling them independently can lead to improved spoofing detection. In particular, it is anticipated that the LE frames capture regions of transition between voiced and unvoiced speech/silence, which in turn are expected to be the regions that would be hardest to model in speech synthesis and voice conversion systems.

3. Experimental Set up

3.1. Voice Activity Detection

In this work we employ a vector quantisation based VAD (VQVAD) [19] for two reasons. Firstly, the VQVAD is self-adaptive and unsupervised in nature and does not require pre-training with voiced or unvoiced frames. Secondly, it does not rely on a fixed threshold and has been shown to be accurate compared to other VADs. For each utterance, the VQVAD generates a voiced GMM based on MFCCs from high energy frames and an unvoiced GMM based on MFCCs from low energy frames. Finally, the voiced-unvoiced distinction is made based on a combination of the log-likelihood ratio between the voiced and unvoiced models and an energy threshold. Specifically frames that lead to a positive log-likelihood ratio and have energy above a preset minimum threshold are assigned as HE frames and all others are assigned as LE frames.

The VQVAD is tuned for the microphone environment with a defined energy threshold. Approximately 20% of the LE frames in the ASVSpoof 2015 corpus was silence, while less than 10% of the LE frames in the SAS corpus were silence. Ideally, we would like the LE frames to comprise of unvoiced frames and frames that capture start and end points of voiced speech, but remove silence frames containing no speech. However, due to the difficulty in separating between these two, we retain all LE frames (unvoiced, voiced transitions and silence). This avoids the risk of potentially removing useful information (other than silence) from other unvoiced segments if they are mistaken to be silence.

3.2. Database

Three databases are employed in our experiments. The Spoofing and Anti-Spoofing (SAS) corpus and the ASVSpoof 2015 corpora are utilized for evaluating logical access scenarios and the BTAS 2016 corpus for evaluating physical access attacks [20].

3.2.1. ASVSpoof 2015 Corpus

This corpus was introduced for the ASVSpoof 2015 challenge [21] and it contains genuine and spoofed speech generated using 3 speech synthesis and 7 voice conversion algorithms. The evaluation data contains both known and unknown attack with the unknown attacks obtained using five additional unseen (in the training and development sets) spoofing algorithm.

3.2.2. Spoofing and anti-spoofing (SAS) Corpus

The SAS corpus [22] consists of genuine, 5 SS and 8 VC based data. Similar to ASVSpoof 2015 training and development dataset contains 5 different spoofing attacks based speech. SAS has the most diverse nature with 8 unseen attacks in the evaluation set. Even though SAS and ASVSpoof15 have the similar training set data, SAS ones have been preprocessed to remove the nonrealistic silence regions which appeared in ASVSpoof 2015 spoofing data. This makes the SAS corpus a more realistic logical access spoofing corpus.

3.2.3. Biometric Theory, Applications and Systems (BTAS 16) Corpus

The BTAS 16 corpus [23] consists of genuine replay, SS replay and VC replay attacks. They were replayed back through normal and high quality conditions. We have used the genuine, SS and VC data for the experiments.

3.3. Front-end

IMFCC and LFCC features (20 dimensions each) were extracted using 20ms frames with 50% overlap and the corresponding 40 dimensional dynamic features comprising of only Δs and $\Delta\Delta s$ were employed as features as in [5] for experiments on both the SAS and the ASVSpoof 2015 corpora. The experiments on the BTAS 2016 employed 60 dimensional feature vectors comprising the static, Δ and $\Delta\Delta$ features. Static features have been more useful in detecting physical access attacks than in detecting logical access attacks [9]. The proposed approach is compared to baseline systems that employ identical features but do not distinguish between HE and LE frames.

3.4. Back-end

A GMM based back-end is employed for both LE and HE frames in the proposed approach, as well as the baseline system. All GMMs were trained using the EM algorithm to obtain the maximum likelihood estimate starting with random initialization. Log-likelihood ratios were computed as

$$LLR(X) = \log P\left(X|\theta_g\right) - \log P(X|\theta_s) \quad (1)$$

where X denotes feature vectors, and θ_g and θ_s denote the GMMs for genuine and spoofed speech, respectively.

In our baseline systems we employ a GMM with 512 mixture components. For the proposed approach, we investigate the use of 512 and 256 mixture GMMs for HE frames and 512, 256 and 128 mixture GMMs for LE frames, given the lower number of LE frames in all databases. Where appropriate we denote the different models as HE256, LE128, etc.

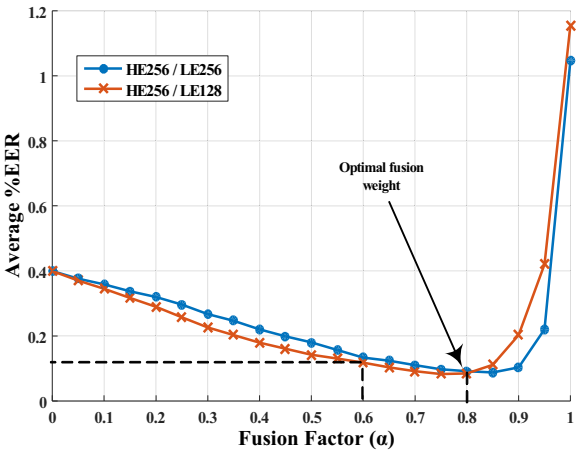

Figure 3: *%EER after score level fusion for different α using LFCC features on SAS development set.*

3.5. Score Level Fusion

In order to combine the HE and LE classifier outputs, a linear regression based score fusion is used, i.e,

$$LLR_{fused} = (1-\alpha)LLR_{HE} + (\alpha)\, LLR_{LE} \quad (2)$$

where LLR_{fused} is the fused log-likelihood score of HE frame log-likelihood ratio, LLR_{HE} and the LE frame log-likelihood ratio, LLR_{LE} and the fusion weight α is determined on the development set.

Both the number of mixtures and the fusion weight are jointly optimized on the development set of the SAS corpus which is comprised of 65% HE frames and 35% LE frames. Specifically, two combinations of the proposed approach, one employing 256 mixtures for both HE and LE frames and the other employing 256 mixtures for HE frames and 128 mixtures for LE frames, are evaluated for a range of different fusion weight values, α and the results are shown in Figure 3.

Based on these results in Figure 3, we choose a system employing 256 mixture GMMs for HE frames and 128 mixture GMMs for LE frames and a fusion weight of 0.8 for all further experiments. It should be noted that the EER of the baseline system is 0.12%, corresponding to $\alpha = 0.6$ in Figure 3. This suggests that the proposed approach places a greater weight on information from the LE frames compared to the baseline system that does not separate HE and LE frames.

4. Evaluation Set Results

Table 2 shows the spoofing detection accuracy obtained on the SAS corpus for the various spoofing attacks included in the evaluation set. In addition to the aforementioned baseline systems, we also compare the proposed approach using LFCCs and IMFCCs to spoofing detection systems employing CQCC features [6] and scattering cepstral coefficients (SCC) [7], which are the best performing front-ends reported in the literature for the SAS corpus. Both of these systems also employ 512 mixture GMMs in the back-end. From these results it can be seen that the proposed approach of separating HE and LE frames outperforms the baseline approach as well as the systems based on CQCCs and SCCs. To the best of the authors' knowledge, this is the best performance on the SAS corpus till date.

Table 1*: Comparison of results for ASVSpoof 2015 (ASV15) and BTAS 16 corpora in terms of Average %EER (K - Known attacks, U-Unknown attacks, All – All attacks).*

		Baseline				**Proposed (HE + LE)**	
		IMFCC [5]	LFCC [5]	CQCC [6, 9]	SCC [7]	IMFCC	LFCC
ASV15	(K)	0.15	0.11	0.05	0.03	0.05	0.05
	(U)	1.86	1.67	0.46	0.80	0.99	0.91
	All	1.01	0.90	0.26	0.42	0.52	0.48
BTAS16	SS	0.03	0.09	0.00	-	0.00	0.00
	VC	0.23	0.16	0.10	-	0.10	0.08

Table 1 presents the evaluation set results from the ASVSpoof 2015 and BTAS 16 corpora. Once again it can be

Table 2*: Results on Evaluation set of SAS corpus with respect to %EER.*

SAS Corpus		Baseline				Proposed (HE + LE)	
	Spoofing Methods	IMFCC (Δ & Δ^2 only)	LFCC (Δ & Δ^2 only)	CQCC [6]	SCC [7]	IMFCC (Δ & Δ^2 only)	LFCC (Δ & Δ^2 only)
Known Attacks	'VC_C1'	0.31	0.34	0.24	0.03	0.30	0.31
	'VC_FS'	0.04	0.04	0.02	0.02	0.03	0.02
	'VC_FESTVOX'	0.12	0.11	0.14	0.02	0.07	0.11
	'SS_LARGE-16k'	0.00	0.00	0.00	0.02	0.00	0.00
	'SS_SMALL-16k'	0.00	0.00	0.00	0.01	0.00	0.00
Unknown Attacks	'VC_KPLS'	0.07	0.20	0.12	0.09	0.03	0.02
	'VC_GMM'	0.14	0.09	0.17	0.02	0.09	0.14
	'VC_EVC'	1.78	0.19	5.02	0.05	1.75	0.04
	'VC_TVC'	2.03	0.29	4.81	0.15	1.98	0.05
	'VC_LSP'	0.02	0.01	0.09	0.04	0.01	0.00
	'SS_LARGE-48k'	0.00	0.00	0.00	0.00	0.00	0.00
	'SS_SMALL-48k'	0.00	0.00	0.00	0.00	0.00	0.00
	'SS_MARY_LARGE'	8.69	11.41	0.61	10.01	2.93	4.50
EER	Known Average	0.09	0.10	0.08	0.02	0.08	0.09
	Unknown Average	1.59	1.52	1.35	1.29	0.85	0.59
	Average	1.02	0.98	0.86	0.80	**0.55**	**0.40**

seen that the proposed approach outperforms the baseline system that does not distinguish between HE and LE frames. The proposed approach does not outperform the SCC and CQCC systems on the ASVspoof15 database. This may be due to the high proportion of nearly silent regions (20% of LE frames), whereby the benefit of retaining all LE frames (avoiding potential loss of information due to misidentification of silence) is outweighed by the cost of including a larger number of silence frames.

It is interesting to note the performance of the proposed approach in detecting the 'SS-Mary' spoofing attack, which has previously been shown to be a particularly hard spoofing method to detect [24]. We believe that this is due to the more highly weighted LE frames capturing discriminative information present at the boundaries of the concatenated speech units.

5. Conclusions

This paper introduces a new method to independently model low and high energy frames in parallel prior to score fusion. This approach places a greater emphasis on low energy frames compared to the standard approach, which in turn is beneficial since these low energy frames encompass the voiced-unvoiced boundaries in speech that are not modelled well by spoofing methods. This approach of separating low and high energy frames using an appropriate VAD and employing independent spoofing models has been validated on all publically available spoofing corpora and the results consistently show that the proposed approach is superior to the standard approach of not distinguishing between low and high energy frames.

6. References

[1] "Biometric Voice Authentication System Nuance", Australia.nuance.com, 2017. [Online]. Available: http://australia.nuance.com/for-business/customer-service-solutions/voice-biometrics/vocalpassword/index.htm.

[2] "Biometric Banking with Fingerprints, Facial and Voice Recognition | USAA", Usaa.com, 2017. [Online]. Available: https://www.usaa.com/inet/pages/enterprise_howto_biometrics_landing_mkt.

[3] 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, Feb. 2015.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

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

[6] M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," *Odyssey*, 2016.

[7] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-End for Anti-Spoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4,pp.632-643, June 2017. doi: 10.1109/JSTSP.2016.2647202.

[8] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: a comparison.," in *INTERSPEECH*, 2015, pp. 2057–2061.

[9] D. Paul, M. Sahidullah, and G. Saha, "Generalization Of Spoofing Countermeasures: A Case Study With ASVSpoof

2015 And BTAS 2016 Corpora," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[10] H. Wu, Y. Wang, and J. Huang, "Identification of reconstructed speech," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, p. 10, 2017.

[11] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, vol. 2, pp. 1303–1306.

[12] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+ noise model," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, vol. 2, pp. 550–553.

[13] E. Godoy, O. Rosec, and T. Chonavel, "Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, May 2012.

[14] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[15] G. P. Kafentzis and Y. Stylianou, "High-resolution sinusoidal modeling of unvoiced speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 4985–4989.

[16] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models.," in *Interspeech*, 2012, pp. 859–862.

[17] K. Tokuda and H. Zen, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5640–5644.

[18] T. G. Csapó, G. Németh, M. Cernak, and P. N. Garner, "Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder," in *Signal Processing Conference (EUSIPCO), 2016 24th European*, 2016, pp. 1338–1342.

[19] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data.," in *ICASSP*, 2013, pp. 7229–7233.

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

[21] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *Training*, vol. 10, no. 15, p. 3750, 2015.

[22] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4440–4444.

[23] P. Korshunov, S. Marcel, H. Muckenhirn, A. Gonçalves, A. S. Mello, R. V. Violato, F. Simoes, M. Neto, M. de Assis Angeloni, J. Stuchi, and others, "Overview of BTAS 2016 speaker anti-spoofing competition," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, 2016, pp. 1–6.

[24] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: the Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, 2017.