Impact of bandwidth and channel variation on presentation attack detection for speaker verification

Héctor Delgado[1], Massimiliano Todisco[1], Nicholas Evans[1], Md Sahidullah[2],
Wei Ming Liu[3], Federico Alegre[3], Tomi Kinnunen[2] and Benoit Fauve[3]
[1]*EURECOM, France* – [2]*University of Eastern Finland, Finland* – [3]*Validsoft Ltd., United Kingdom*
{delgado,todisco,evans}@eurecom.fr, {sahid,tkinnu}@cs.uef.fi,
{jasmin.liu,federico.alegre,benoit.fauve}@validsoft.com

***Abstract*—Vulnerabilities to presentation attacks can undermine confidence in automatic speaker verification (ASV) technology. While efforts to develop countermeasures, known as presentation attack detection (PAD) systems, are now under way, the majority of past work has been performed with high-quality speech data. Many practical ASV applications are narrowband and encompass various coding and other channel effects. PAD performance is largely untested in such scenarios. This paper reports an assessment of the impact of bandwidth and channel variation on PAD performance. Assessments using two current PAD solutions and two standard databases show that they provoke significant degradations in performance. Encouragingly, relative performance improvements of 98% can nonetheless be achieved through feature optimisation. This performance gain is achieved by optimising the spectro-temporal decomposition in the feature extraction process to compensate for narrowband speech. However, compensating for channel variation is considerably more challenging.**

***Index Terms*—presentation attack detection, speaker verification, bandwidth and channel variation**

I. Introduction

While automatic speaker verification (ASV) [1]–[3] offers a convenient, reliable and cost-effective approach to person authentication, vulnerabilities to presentation attacks [4], previously referred to as spoofing, can undermine confidence and form a barrier to exploitation. By masquerading as enrolled clients, fraudsters can mount attacks to gain unauthorised access to systems or services protected by biometrics technology.

Presentation attacks in the context of ASV can be performed with impersonation, speech synthesis, voice conversion and replay [5]. While the study of impersonation has received attention, e.g. [6], replay, speech synthesis and voice conversion are assumed to pose the greatest threat [7]. Speech synthesis and voice conversion presentation attacks combine suitable training or adaptation data with sophisticated algorithms which generate voice samples whose spectral characteristics resemble those of a given target speaker. In contrast, replay spoofing attacks require neither specialist expertise nor equipment and can hence be mounted by the lay person with relative ease. Replay attacks involve the re-presentation to an ASV system of another person's speech which is captured beforehand, possibly surreptitiously, for instance during an access attempt.

The study of presentation attack detection (PAD) for ASV is now an established area of research [7]. The first competitive evaluation, namely the ASV spoofing and countermeasures (ASVspoof) challenge [8], was held in 2015. It promoted the development PAD solutions to protect ASV from voice conversion and speech synthesis attacks.

Since the first ASVspoof 2015 evaluation, the community has started to consider a number of more practical aspects of PAD. Some recent work has explored the impact of additive noise on reliability [9], [10] and the benefit of speech enhancement and multi-condition training as a means of improving robustness [9], [11].

Other likely influences on PAD performance, e.g. bandwidth and channel variability, have received comparatively little attention to date [12], [13]. Given the prevalence of ASV technology in telephony applications were bandwidth is typically low and where coding, packet loss and other non-linear effects have potential to degrade performance, these aspect require attention. However, the ASVspoof 2015 database contains high quality, high bandwidth recordings. The RedDots Replayed database [14] which was generated from the text-dependent ASV RedDots database [15], was introduced recently to support the development PAD solutions for replay presentation attacks. While exhibiting variation in terms of recording devices and environmental conditions, and hence representing a greater degree of practical, real-life variability, it still contains wideband audio (16kHz).

The work reported in this paper has accordingly sought to investigate the impact of bandwidth and channel variation on PAD reliability for ASV. The work was performed with bandwidth-limited and coded versions of the ASVspoof 2015 and RedDots Replayed databases (covering 3 different types of presentation attacks, namely speech synthesis, voice conversion and replay), generated through band-pass filtering, downsampling and coding. The work was performed with two PAD solutions, namely linear frequency cepstral coefficients [16] and constant Q cepstral coefficients [17], [18], both of which achieve competitive performance for the ASVspoof 2015 database with a relatively simple back-end classifier. It is stressed that the objective of the work reported here is to assess the impact on PAD reliability of bandwidth and channel variation. While an issue of undoubtable importance, the work is NOT concerned with generalisation.

II. Presentation attack databases

The work reported in this paper was performed using two publicly available databases.

A. ASVspoof 2015

The ASVspoof initiative [8] was the first community-led effort to collect a common database to support research in spoofing and countermeasures. The ASVspoof 2015 database contains a mix of *bona fide* (genuine speech without attack) and spoofed speech. All bona fide speech data is sampled at 16kHz and was recorded in a semi-anechoic chamber with a solid floor [8]. Spoofed speech is generated with 10 different speech synthesis and voice conversion algorithms. In order to promote generalised PAD systems, only 5 of these were used to generate training and development subsets whereas an evaluation subset was generated with the full 10. In this paper, the development set containing genuine and spoofed speech using 5 different attacks is used. Table I shows database statistics. Full details of the ASVspoof 2015 database and example PAD results are available in [8].

TABLE I
Statistics of the ASVspoof 2015 database: number of speakers (M=male, F=female), and number of genuine and spoofed trials.

Partition	#Speakers (M / F)	#Genuine trials	#Spoofed trials
Training	10 / 15	3750	12625
Development	15 / 20	3497	49875

B. RedDots Replayed

The **RedDots Replayed** database [14] was designed to support the development of PAD solutions for replay attacks in diverse recording and playback environments. RedDots Replayed is based upon the re-recording of the original RedDots database [15] (part 01, male speakers) which contains speech data comprising 10 common passphrases recorded in a number of acoustic conditions using mobile devices with a sampling rate of 16kHz. Replayed speech is generated with one of 16 different recording devices, 15 different playback devices and various different acoustic conditions, including both controlled and more variable (unpredictable) conditions. Controlled condition recordings are made in a quiet office/-room whereas variable condition recordings are made in noisier environments. A training subset contains only controlled condition recordings whereas an evaluation subset contains both controlled and variable condition recordings. Table II shows database statistics. Full details of the RedDots replayed database and example presentation attack detection results are available in [14]. A subset of the RedDots Replayed database is also used in the ASVspoof 2017 challenge[1] data [19], [20].

[1]http://www.asvspoof.org/

TABLE II
Statistics of the RedDots Replayed database: number of speakers (male), and number of genuine and spoofed trials.

Partition	#Speakers	#Genuine trials	#Spoofed trials
Training	10	1508	9232
Development	39	2346	16067

C. Bandwidth reduction and channel simulation

PAD performance was assessed with different versions of each database: (i) the original full-band versions; (ii) bandwidth-reduced versions, and (iii) versions with additional channel variation simulated with the Idiap acoustic simulator software[2].

Bandwidth reduction involves downsampling from 16kHz to 8kHz. ITU G.151[3] compliant bandpass filtering is applied with a gain of -3dB at the passband edges of 300Hz and 3400Hz. The original and bandwidth-reduced versions are referred to from hereon as **wideband** (WB) and **narrowband** (NB).

Codec simulations employ a common ITU G.712[4] compliant bandpass filter. This is combined with a-law coding[5] at a rate of 64kbit/s for landline telephony and with an adaptive multi-rate narrowband (AMR-NB) codec[6] at a rate of 7kbit/s for cellular telephony. These two scenarios are referred to as **landline** (L) and **cellular** (C), respectively. Figure 1 illustrates the distortion in the long-term average spectrum for landline and cellular coded signals compared to the original narrowband signal for an arbitrary speech utterance from the ASVspoof 2015 database. These spectra were obtained with the constant Q transform (CQT, see Section III). In addition to broad attenuation, the plots illustrates substantial spectral distortion, especially at lower and higher frequencies. The distortion is particularly pronounced for the cellular-coded signal.

III. Presentation attack detection

The work was performed with two different PAD systems. A backend Gaussian mixture model (GMM) classifier with two classes, one for fona fide speech and one for spoofed speech is common to both systems. Models are learned using bona fide and spoofed data from their respective training subsets and with an expectation maximisation algorithm. According to independent results, e.g. [16], [18], [21], such a simple classifier often provides competitive or even better performance compared to other, more sophisticated algorithms. The score for a given trial is computed as the log-likelihood ratio of the test speech sample given the two GMMs for bona fide and spoofed speech. The frontends are described below. Neither employs voice activity detection.

[2]http://github.com/idiap/acoustic-simulator

[3]https://www.itu.int/rec/T-REC-G.151-198811-W/en, accessed: 2017-08-07

[4]https://www.itu.int/rec/T-REC-G.712/en, accessed: 2017-08-07

[5]https://www.itu.int/rec/T-REC-G.711-198811-I/en, accessed: 2017-08-07

[6]https://www.itu.int/rec/T-REC-G.711-198811-I/en, accessed: 2017-08-07

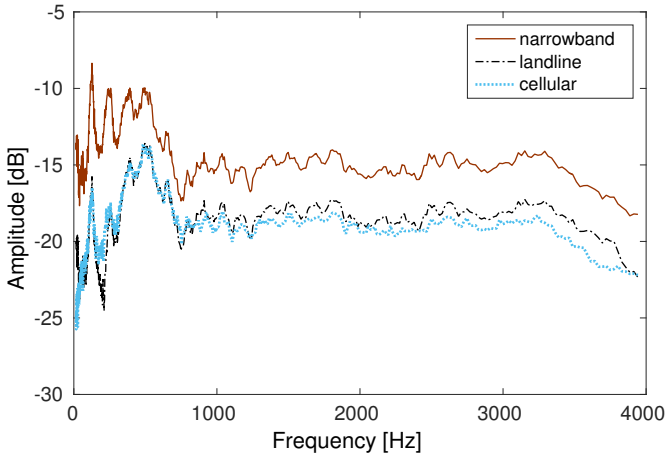

Fig. 1. Average long-term CQT spectra for the utterance 'He's worked for several years in the United States' for narrowband, landline and cellular channels.

The **linear-frequency cepstral coefficient** frontend is the best performing system from [16]. The energy outputs of a uniformly-spaced, triangular filterbank are processed by the discrete cosine transform (DCT) to derive cepstral coefficients using an analysis window of 20ms with a 10ms shift. Since LFCC features are computed with linearly-spaced filters, the frequency resolution is explicitly related to the number filters. Increasing the number improves the frequency resolution and captures more detailed spectral characteristics. While the original work [16] used 20 filters, use of a greater number was found to improve performance. For work reported here, the number of filters is optimised first for WB and then for NB data.

Constant Q cepstral coefficients. The second front-end involves constant Q cepstral coefficients (CQCCs) [17], [18] which combine the constant Q transform (CQT) [22] with standard cepstral analysis. In contrast to Fourier techniques, the centre/bin frequencies of the CQT scale are geometrically distributed [23]. The centre frequency f_k for the k-th frequency bin is given by

$$f_k = f_{min} 2^{\frac{k-1}{B}} \quad (1)$$

where f_{min} is the minimum frequency considered and B is the number of bins per octave. Higher values of B provide greater frequency resolution but reduced time resolution, while lower values of B provide greater time resolution but smaller frequency resolution. B thus determines the trade-off between frequency and time resolutions and is a major optimisation parameter for CQT-based analysis. Note that the CQCC analysis window length and shift is effectively variable in order to maintain a constant Q factor (trade-off between centre frequency and filter width) across frequency bins. Full details of CQCC extraction are described in [18].

IV. Experimental work

This section reports an assessment of bandwidth and channel variation impacts on PAD performance. All experiments were performed with the standard protocols in [8], [14] (see Section II). Assessments are based on the threshold-free equal error rate (EER_{pad}) metric for a bona fide/presentation attack discrimination task. EER_{pad} is the operating point where the attack presentation classification error rate, APCER (equivalent to the false alarm rate, FAR, in binary classification tasks), and the bona-fide presentation classification error rate, BPCER (equivalent to miss rate in binary classification tasks), are equal. Shown first are baseline experiments using the original high-quality WB versions of the ASVspoof 2015 development set (in the following referred to as ASVspoof) and RedDots Replayed database. The use of the ASVspoof development set alone avoids any influence of results on presentation attacks for which no training data is available; **this paper is not concerned with generalisation aspects**. Then, the adopted methodology is summarised as follows:

- Baseline experiments using the original high-quality WB databases were performed.
- Identical experiments using NB versions of the same databases were performed to evaluate performance for bandwidth-reduced audio.
- Feature extraction configurations are optimised to improve performance for bandwidth-reduced audio.
- A final set of experiments evaluate the robustness of optimised PAD solutions in the face of additional speech coding.

A. Wideband baseline

Baseline results for LFCC and CQCC features and the original WB databases (no downsampling nor channel simulation) are presented in Table III ("Wideband" rows). LFCCs include 20 delta (D) and 20 acceleration (A) coefficients [16] computed using 30 filters while CQCCs include 20 A coefficients [18]. These configurations were optimised for the ASVspoof database. Error rates for LFCC features are twice those of CQCC features. Error rates for the RedDots Replayed database are markedly higher than for the ASVspoof database, albeit that these results were generated using un-optimised feature configurations.

B. Bandwidth reduction

Table III ("Narrowband" rows) shows results for the NB versions of ASVspoof and RedDots Replayed databases. Results are shown for both LFCC and CQCC features using different combinations of static (S), delta (D) and acceleration (A) coefficients. Results in Table III show that, for the ASVspoof database, performance is significantly degraded for both LFCC and CQCC features. For LFCC features, the EER_{pad} increases from 0.11% to 1.64% whereas that for CQCC features increases from 0.05% to 9.92%. In addition, for CQCC, SD configuration further reduces the error rate of A configuration further down to 5.64%.

For the RedDots Replayed database, performance for LFCC features degrades from 6.18% to 8.12%. For CQCC features, results improve, with the EER_{pad} dropping from 3.27% to 2.07%. Our analysis suggests that this is because the salient

TABLE III
PERFORMANCE OF LFCC AND CQCC PAD SYSTEMS IN TERMS OF EER_{pad} (%) FOR ASVSPOOF DEVELOPMENT AND REDDOTS REPLAYED DATABASES FOR **WB** AND **NB** DATA. PAD SYSTEMS WERE NOT OPTIMIZED FOR NB DATA. S=STATIC, D=DELTA, A=ACCELERATION.

	Feature		ASVspoof 2015	RedDots Replayed
Wideband (16 kHz)	LFCC	DA	0.11	6.18
	CQCC	A	0.05	3.27
Narrowband (8 kHz)	LFCC	S	6.60	13.30
		D	3.38	9.02
		A	4.06	8.24
		SD	3.72	10.27
		SA	3.17	9.56
		DA	**1.64**	**8.12**
		SDA	2.27	8.59
	CQCC	S	10.39	7.13
		D	10.93	3.18
		A	**9.92**	**2.07**
		SD	5.64	4.05
		SA	5.90	4.18
		DA	8.97	2.14
		SDA	5.71	2.88

information for replay detection is contained within low frequencies for which CQCC features have better resolution. The same behaviour is not observed for LFCC features, however. This is because LFCC features may lack sufficient resolution at low frequencies to capture the same information captured by CQCC features.

While it is not entirely surprising that different features are best for the ASVspoof and RedDots Replayed databases – they contain presentation attacks of a different nature – performance is sensitive to the particular configuration. Whereas DA and A combinations give the best performance for WB ASVspoof data for LFCC and CQCC features respectively, DA and SD combinations give the best performance for NB data. Performance for the RedDots Replayed database is more consistent with DA and A configurations again giving the best performance.

C. Feature optimisation

Reported now are results for optimised LFCC and CQCC features for NB data. For LFCC features, optimisation is performed by varying the number of filters. The dimensionality of static features is fixed by considering first 20 coefficients after the DCT. Table IV reports results for ASVspoof and RedDots Replayed databases where the number of filters is varied between 20 and 80. For the ASVspoof database, performance is improved for a higher number of filters. The best performance is obtained with 70 filters and dynamic coefficients (DA). However, for the RedDots Replayed database, the optimal number of filters is 30 while performance degrades for higher numbers.

Table V shows optimisation results for CQCC features. Performance is illustrated for different combinations of S, D and A coefficients and as a function of the number of

TABLE IV
OPTIMISATION OF NUMBER OF FILTERS FOR **LFCC** FEATURES FOR **NB** ASVSPOOF DEVELOPMENT AND REDDOTS DATABASES IN TERMS OF EER_{pad} (%) FOR DIFFERENT CONFIGURATIONS OF STATIC (S), DELTA (D) AND ACCELERATION (A) COEFFICIENTS.

		20	30	40	50	60	70	80
ASVspoof 2015	S	**5.74**	6.60	6.19	6.12	6.34	6.45	6.52
	D	4.48	3.38	3.28	**3.19**	3.21	3.21	3.25
	A	5.21	4.06	4.05	4.05	3.94	**3.91**	4.04
	SD	**3.48**	3.72	3.62	3.67	3.64	3.65	3.49
	SA	3.27	3.17	**3.04**	3.21	3.13	3.16	3.08
	DA	2.10	1.64	1.67	1.49	1.50	**1.44**	1.55
	SDA	2.34	2.27	2.18	2.21	2.13	2.16	**2.06**
RedDots Replayed	S	13.71	**13.30**	13.51	13.97	14.45	15.18	15.30
	D	9.06	**9.02**	9.51	9.66	10.14	10.05	10.60
	A	**8.13**	8.24	8.48	8.52	8.97	9.15	9.26
	SD	10.67	**10.27**	10.87	11.64	11.61	11.72	11.74
	SA	9.97	**9.56**	10.14	10.38	10.72	11.08	11.13
	DA	8.40	**8.12**	8.40	9.08	8.72	9.04	9.40
	SDA	9.11	**8.59**	9.63	9.65	10.17	10.57	10.53

TABLE V
OPTIMISATION OF THE NUMBER OF FREQUENCY BINS PER OCTAVE B FOR **CQCC** FEATURES FOR **NB** ASVSPOOF AND REDDOTS REPLAYED DATABASES IN TERMS OF EER_{pad} (%) FOR DIFFERENT CONFIGURATIONS OF STATIC (S), DELTA (D), AND ACCELERATION (A) COEFFICIENTS.

	B	192	96	48	24	12	6
ASVspoof 2015	S	17.23	10.39	5.25	2.95	**1.93**	3.06
	D	16.01	10.93	7.11	5.64	**4.53**	6.27
	A	14.73	9.92	8.08	6.40	**4.88**	8.69
	SD	10.97	5.64	2.72	1.00	**0.28**	0.37
	SA	10.45	5.90	3.35	1.05	**0.17**	0.31
	DA	13.29	8.97	6.25	4.44	**3.54**	5.70
	SDA	10.30	5.71	2.60	0.84	**0.16**	0.27
RedDots Replayed	S	**6.57**	7.13	8.82	10.06	9.68	
	D	3.50	**3.18**	3.46	7.55	11.68	
	A	2.50	**2.07**	3.20	4.65	9.21	
	SD	**3.88**	4.05	5.43	7.20	7.74	-
	SA	**3.85**	4.18	5.63	7.30	8.15	
	DA	2.73	**2.14**	2.6	4.82	11.05	
	SDA	**2.86**	2.88	3.86	6.22	8.44	

bins per octave B involved in the CQT computation. The combination of SDA coefficients gives the best performance for the ASVspoof database (0.16% EER_{pad} for B=12) whereas A coefficients alone give the more consistent performance for the RedDots database (2.07% EER_{pad} for B=96). In terms of general trends, smaller values of B give better performance for the ASVspoof database whereas larger values of B give better performance for the RedDots database. This would suggest that the detection of voice conversion and speech synthesis attacks requires a spectro-temporal analysis with higher time resolution. Conversely, the reliable detection of replay attacks requires a higher frequency resolution.

D. Channel simulation

For experiments described above, PAD algorithms were optimised for a 'generic' telephony scenario through the downsampling of original WB data to NB data. Experiments

reported here focus on the evaluation of PAD systems on more challenging data with simulated landline (L) and cellular (C) channel variation. Results are presented in Table VI for the optimised PAD systems corresponding to Tables IV and V. LFCC features have dynamic coefficients (DA) computed using 70 filters for the ASVspoof database. For the RedDots Replayed database, features are the same, except for 30 filters. Performance degrades significantly for both landline and cellular scenarios, more so for the ASVspoof database than for the RedDots Replayed database.

TABLE VI
PERFORMANCE OF OPTIMUM CONFIGURATIONS FOUND IN SECTION IV-C APPLIED TO THE ASVSPOOF AND REDDOTS REPLAYED DATABASES WITH **SIMULATED CELLULAR (C) AND LANDLINE (L) CHANNELS** (RESULTS FOR NARROWBAND (NB) ALSO INCLUDED FOR COMPARISON).

	ASVspoof			RedDots Replayed		
	NB	L	C	NB	L	C
LFCC	1.44	6.05	**11.09**	8.12	8.38	**10.14**
CQCC	**0.16**	**1.86**	12.96	**2.07**	**3.10**	12.32

CQCC features involve the full SDA configuration with B=16 frequency bins per octave for the ASVspoof database and A coefficients with B=96 frequency bins per octave for the RedDots Replayed database. Performance again degrades significantly for both landline and cellular scenarios and, again, much more for the latter. The relative degradation for CQCC features in the case of the cellular scenario is significantly greater than for LFCC features. This could indicate that, despite seemingly better performance for matched conditions, CQCC features are more sensitive to channel variation than LFCC features. Given that both landline and cellular scenarios share the same bandpass filtering, the degradation stems from the use of different codecs. The AMR-NB codec has a high compression rate of 7kbits/s. This degradation in performance most likely stems from aggressive compression and the consequential loss of frequency components which are crucial for presentation attack detection.

To further illustrate PAD performance degradation due to codec effects, Figure 2 shows DET plots of the CQCC PAD system for generic narrowband, landline and cellular scenarios on the RedDots replayed database (replay attacks). PAD on narrowband data is more accurate than on landline data for a wide range of operation points. PAD performance on cellular data is importantly degraded for the complete range of operation points.

V. CONCLUSIONS

This paper reports an investigation of bandwidth and channel variation on the reliability of presentation attack detection (PAD) for automatic speaker verification. Experiments were performed using two common databases of spoofed speech, namely ASVspoof 2015 and RedDots Replayed which, together, contain a variety of different presentation attacks. Results show that the performance of two state-of-the-art PAD solutions optimised for WB speech degrades significantly

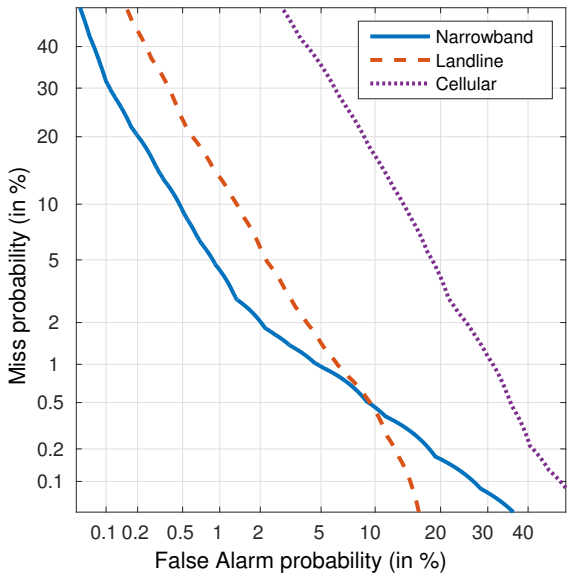

Fig. 2. DET plots for narrowband, landline and cellular scenarios on the RedDots replayed database.

when applied to NB speech, while PAD optimisation can improve performance. A higher frequency resolution might be needed for the detection of replay attacks whereas higher time resolution is need for the detection of voice conversion and speech synthesis attacks. In the face of channel variation, performance again degrades significantly. These findings show the need for new, common databases of spoofed speech which incorporate channel variation in addition to new research in channel compensation for PAD.

ACKNOWLEDGMENT

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

REFERENCES

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.
[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.
[3] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.
[4] "ISO/IEC 30107-3: Information technology – biometric presentation attack detection," International Organization for Standardization, Standard, 2016.
[5] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. INTERSPEECH*, 2013, pp. 925–929.
[6] R. Hautamki, T. Kinnunen, V. Hautamki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13 – 31, 2015.

[7] 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.

[8] 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*, vol. PP, no. 99, pp. 1–1, 2017.

[9] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83 – 97, 2016.

[10] X. Tian, Z. Wu, X. Xiao, E.-S. Chng, and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions," in *Proc. INTERSPEECH*, 2016, pp. 1715–1719.

[11] H. Yu, A. K. Sarkar, D. A. L. Thomsen, Z.-H. Tan, Z. Ma, and J. Guo, "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *Proc. International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016.

[12] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[13] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, 2011, pp. 1–8.

[14] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. González-Hautamäki, D. Thomsen, A. K. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K.-A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICASSP*, 2017.

[15] K.-A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2996–3000.

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

[17] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," in *Odysey - the Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016.

[18] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, 2017.

[19] T. Kinnunen, N. Evans, J. Yamagishi, K.-A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," http://www.asvspoof.org/data2017/asvspoof_2017_evalplan_v0.pdf, 2017.

[20] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.-A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017.

[21] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. INTERSPEECH*, 2015, pp. 2062–2066.

[22] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proc. ICASSP*, vol. 3, Apr 1978, pp. 375–378.

[23] R. E. Radocy and J. D. Boyle, *Psychological foundations of musical behavior*. C. C. Thomas, 1979.