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

Audio Replay Attack Detection Using High-Frequency Features

Marcin Witkowski, Stanisław Kacprzak, Piotr Żelasko, Konrad Kowalczyk, Jakub Gałka

AGH University of Science and Technology
Department of Electronics, Kraków, Poland
{witkow|skacprza|pzelasko|konrad.kowalczyk|jgalka}@agh.edu.pl

Abstract

This paper presents our contribution to the ASVspoof 2017 Challenge. It addresses a replay spoofing attack against a speaker recognition system by detecting that the analysed signal has passed through multiple analogue-to-digital (AD) conversions. Specifically, we show that most of the cues that enable to detect the replay attacks can be found in the high-frequency band of the replayed recordings. The described anti-spoofing countermeasures are based on (1) modelling the subband spectrum and (2) using the proposed features derived from the linear prediction (LP) analysis. The results of the investigated methods show a significant improvement in comparison to the baseline system of the ASVspoof 2017 Challenge. A relative equal error rate (EER) reduction by 70% was achieved for the development set and a reduction by 30% was obtained for the evaluation set.

Index Terms: anti-spoofing, replay detection, playback detection, speaker recognition

1. Introduction

The efficacy of recent Automatic Speaker Verification (ASV) systems in terms of determining whether the voice of a speaker matches the claimed identity is generally high [1–3]. Considering the maturity of voice biometrics technology, the security of such systems must be guaranteed also. Development of methods that increase the robustness of speaker recognition to a variety of attacks is considered a pre-requisite to its widespread commercial applications.

Spoofing attack is an act of deceiving a biometric system in order to obtain positive verification status given the claimed (attacked) identity. It is usually performed as an attack at a microphone or telecommunication level. Wu et al. identify four main spoofing attack types: impersonation, replay, speech synthesis, and voice conversion [4]. The ASVspoof 2017 Challenge addresses the problem of replay spoofing detection [5]. This kind of spoofing is exemplified by a scenario in which the attacker records the voice of a target speaker and later plays it back in order to deceive a speaker verification system, as presented in Figure 1. As stated in [6], these types of attacks are the most frequent and likely to occur since they do not require major expertise or equipment. The vulnerability of speaker verification systems to replay attacks has been reported e.g. in [7–9].

In [8–10] authors describe playback audio detectors which successfully detect spoofing by comparing a new recording with the previously acquired ones, however this kind of countermeasures rely on the assumption that the original recording is known at the time of attack. Detection of far-field recording and loudspeaker playback has been reported in [11] and the algorithm that identifies an acoustic channel artefacts has been presented in [12]. The authors show that cues for distinguishing genuine and spoofed recordings are present in their amplitude-frequency characteristics.

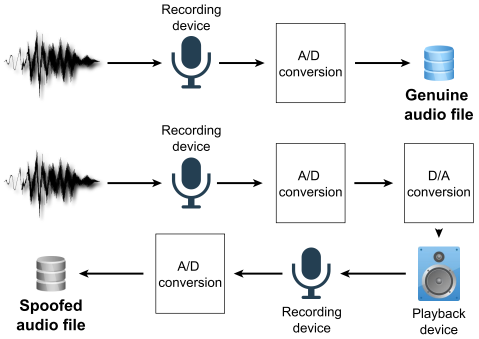

Figure 1: *Genuine (upper) and spoof (lower) file generation scenarios.*

In this paper, we argue that the replay attacks indeed modify the amplitude-frequency characteristics of the audio signal. Specifically, we show that most of the cues can be found in a high-frequency sub-band. We propose and evaluate several replay spoofing countermeasures based on the detection of the observed phenomena and support their potential by the detection accuracy improvement in comparison to the baseline system of the ASVspoof 2017 challenge.

In Section 2, we present a short description of the applied methods and the reasoning we followed in the development of the proposed countermeasures. The data and the experimental set-up are described in Section 3. Section 4 presents the evaluation of the obtained results, followed by conclusions presented in Section 5.

2. Method

In order to construct an informative and discriminative set of features, we have identified three main sources of factors that affect the audio signal in the replay spoofing scenario, namely the playback device, the recording device and the acoustic environment where the recording takes place.

The playback devices are equipped with loudspeakers, which typically have a non-flat magnitude frequency response acting as bandpass filters with non-regular oscillations in the passband [13]. The recording device induces similar effects on the signal. A digital recording device also has an analogue-to-digital converter (ADC) and an associated low-pass anti-aliasing filter with a specified cut-off frequency. Every digital recording is subject to the anti-aliasing filtering, however, in case of a spoofed recording the speech signal undergoes anti-aliasing filtering at least twice. These filters induce modifications (imperfections) near the Nyquist frequency. Finally, the room acoustics where the recording is taking place also has an

http://dx.doi.org/10.21437/Interspeech.2017-776

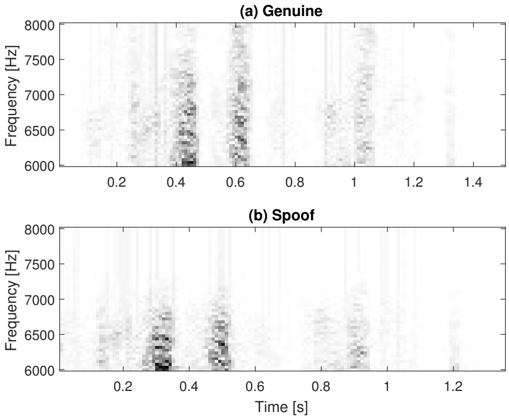

Figure 2: *Spectrum of a genuine (a) and spoofed (b) audio files from the training dataset in the sub-band of 6–8 kHz.*

effect on the recorded signal - most notably - reverberation [14].

Although low-frequency effects caused by the usage of loudspeakers in considered spoofing could be notable, we decided to focus in this work on high-frequency spectral features which capture the effects introduced by analogue-to-digital conversion. In particular, we investigate how the different features computed using high-frequency sub-bands, where anti-aliasing artefacts due to multiple AD conversions occur, contribute to spoofing detection. As an initial investigation, we looked into the differences between the genuine and spoof recordings near the Nyquist frequency. Example spectrograms of genuine and spoof files with the same semantic content for the sampling frequency $f_s = 16$ kHz [1] are shown in Figure 2. Strong low-pass filtering at around 7.25 kHz cut-off frequency and the temporal spectrum scattering caused by reverberation are visible in the spectrogram for spoofed case. In order to identify the appropriate frequency range of interest, we compared the equal error rates obtained using a 2-class Gaussian Mixture Model (GMM) log-likelihood ratio (LLR) classifier based on cepstral and CQCC features extracted for several frequency ranges. We considered the frequency bands with the lower frequency ranging from 1 to 7.5 kHz, while the highest frequency was kept constant at the Nyquist frequency, i.e. at 8 kHz. The EER results obtained using ASVspoof 2017 development dataset depicted in Figure 3 indicate that setting the lower frequency bound in the range between 4 and 6 kHz results in the smallest error, with a minimum at 6 kHz for cepstrum-based features. For narrower frequency bands, i.e. where the lowest frequency is above 6 kHz, a rapid increase in EER results is observed. Consequently, in the following we chose the $4 - 8$ kHz and $6 - 8$ kHz frequency ranges for the selected set of features.

2.1. Features

Based on initial observations of the spectra and EERs, we selected to investigate the frequency features which analyse high frequency content, besides standard broadband features.

2.1.1. Standard broadband features

CQCC: Constant Q Cepstral Coefficients, which are obtained from the Constant Q Transform [15] of a signal, followed by a uniform resampling and a Discrete Cosine Transform (DCT) [16]. These features were chosen as baseline features for the challenge [5].

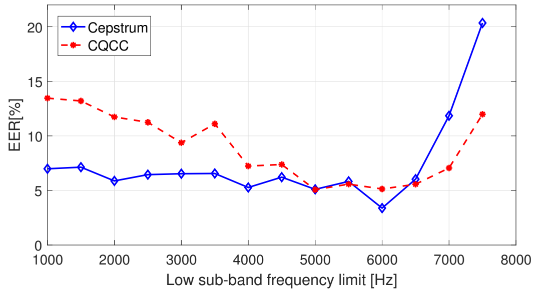

Figure 3: *EER's computed using Cepstrum and baseline CQCC with frequency range low limit presented at horizontal axis.*

Cepstrum: These features are computed as a logarithm of the power of the short-time Fourier spectrum, followed by the DCT applied per frame [17]. Usually, a number of coefficients returned by DCT is limited to 30 or less. Sub-band analysis is performed prior to DCT computation by limiting the number of frequency bins within a spectrum to a specific range.

MFCC: Mel-Frequency Cepstral Coefficients are the most common features used in speech analysis. MFCC is based on cepstral coefficients computed as a logarithm of energies obtained from filtering the signal using a bank of triangular bandpass filters on the mel-frequency scale [18]. The width of subsequent bandpass filters is increasing with frequency. We perform sub-band analysis using outputs of the selected filter banks with central frequencies from the analysed frequency range.

2.1.2. *Proposed features for high frequency analysis*

IMFCC: These features were computed similarly to the MFCCs, but the sequence of filters was inverted in the frequency domain, i.e. high frequencies were represented in more detail. Some advantages of these features in context of spoofing attack detection have been described in [19].

LPCC: Linear Prediction Cepstral Coefficients have been used as one of the common features in speech parametrisation. These features are also assumed here to represent a generalized spectral envelope of an anti-aliasing filter. Linear Prediction coefficients (LPC) are the low order Finite Impulse Response (FIR) filter coefficients that approximate a spectral envelope of an input signal. LPCCs are defined as cepstrum computed from LPC coefficients.

LPCCres: Linear predictive model allows for a decomposition of the speech signal into the linear part that can be predicted using LPC coefficients and the remaining residual signal [20]. Specifically, the residual signal contains all relevant components that are not modelled by linear prediction up to the selected order. We assume that spoofing artefacts are present in a higher frequency region of the residual signal, near the Nyquist frequency. The sub-band residual was modelled with cepstrum and it was subsequently used as a feature which characterizes the remaining components in the microphone signal. This include the detailed fluctuations of the microphone signal such as transients or changes due to multiple AD and DA conversions. Finally, LPCCres features combine both LPCC features concatenated with sub-band cepstrum of a residual. Note that the features based on the residual signal have also been

[1] Respectively files T_1000003.wav and T_1001511.wav from the training set were used as genuine and spoof examples.

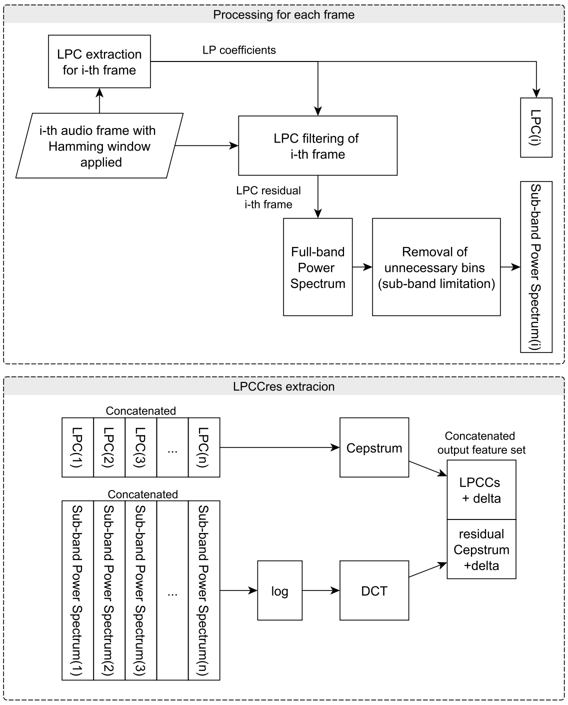
Processing for each frame
LPC extraction for i-th frame
LP coefficients
i-th audio frame with Hamming window applied
LPC filtering of i-th frame
LPC(i)
LPC residual i-th frame
Full-band Power Spectrum
Removal of unnecessary bins (sub-band limitation)
Sub-band Power Spectrum(i)
LPCCres extracion
Concatenated
LPC(1)
LPC(2)
LPC(3)
...
LPC(n)
Cepstrum
Concatenated output feature set
LPCCs + delta
Concatenated
Sub-band Power Spectrum(1)
Sub-band Power Spectrum(2)
Sub-band Power Spectrum(3)
...
Sub-band Power Spectrum(n)
log
DCT
residual Cepstrum +delta

Figure 4: *Extraction of LPCCres features.*

used in the previous anti-spoofing challenge [21,22]. The block diagram of the LPCCres processing is shown in Figure 4.

2.2. Classifier

The spoofing detection can be formulated as a binary classification problem. In this work Gaussian mixture model was used as a back-end classifier, obtained by fitting two separate GMMs to the genuine and spoofed recordings using Expectation-Maximization (EM) training. Classification score was computed as a log-likelihood ratio

$$\mathcal{LLR} = log(\mathcal{L}_{genuine}) - log(\mathcal{L}_{spoof}), \qquad (1)$$

where $\mathcal{L}_{genuine}$ and $\mathcal{L}_{spoof}$ are the test sample likelihoods given the genuine and spoof GMMs respectively. Collected development and evaluation scores were used to estimate the EER, which was the only criterion used to rank the systems in the ASVspoof 2017 Challenge.

3. Performed experiments

3.1. Audio database

In this study the database from ASVspoof 2017 was used as the only data source. The database was created as a subset of processed recordings from RedDots project [23]. It is composed of audio files with 16-bit resolution and 16 kHz sampling rate. For the challenge, the organisers provided three subsets of the database - training, development and evaluation sets, which contained 3016, 1710, and 13306[2] files respectively. Training and development sets were published with *spoof* and *genuine* labels for system preparation. Evaluation contained non-labelled files and was used for system assessment. In the development phase back-end classifiers were trained with the training set, while the development set was used for testing only. For the purpose of the final evaluation, we trained the classifiers using both the training and the development sets. In the entire research the *common condition* described in [5] was followed, i.e. no external data were used in training or adaptation of the presented classifiers.

3.2. Parameterisation of applied features and classifier

For Cepstrum, MFCC, LPCC and IMFCC features, a common framing was used with 25ms frames and 10ms overlap between successive frames, while for LPCCres the frame length was extended to 50ms as it led to higher EERs. Note that CQCCs use varying frame length as described in [16].

For each frame and feature type, the extracted features were: the 0^{th}, 19 static and 19 delta features. If due to sub-band limitation, the number of static coefficients was smaller than 19, the minimal value was chosen automatically.

For both MFCC and IMFCC, 60 filters that cover fullband have been designed, and a triangular filter was selected for further processing if its central frequency belonged to an analysed sub-band. MFCC and IMFCC was computed using the Rastamat toolbox [24]. In LPCC extraction 34^{th}-order filters were approximated using Levinson-Durbin recursion for each frame in full-band frequency range. Note that no sub-band limiting was applied for LPCC (separate and in fusion within LPCCres), and in general the gain parameter was not used in this study. Recursion transformation of LPC coefficients into cepstral coefficients was performed for final feature representation.

To enhance the resolution of CQCC we increased both the default 96 bins-per-octave and 16 as a number of uniform samples in the first octave to 256. To this end, the implementation provided with the baseline system in the challenge was used, as well as its implementation of sub-band limitation.

In all experiments a 512-component GMM with a diagonal covariance matrix was used as a model for both spoof and genuine classes, as we focused on comparison of different features. The MSR Identity Toolbox [25] implementation of the EM GMM training and scoring was used in this research.

3.3. Experiments with other classifiers

In the speaker recognition domain, the GMM and Universal Background Model (UBM) approaches have been outperformed in the recent years by the i-vector framework [3,26,27]. Similarly, deep neural networks (DNN) have been shown to provide state-of-the-art performance in several speech technology domains [28–30]. However, those frameworks typically require large amount of training data - often thousands of hours of recordings [29,30]. We investigated the viability of these approaches given the limited amount of training data in the challenge, which however did not lead to improving the results obtained for the GMM classifier. During these experiments, we observed that both i-vector and DNN [3] models tend to over-fit the training data, and in consequence they did not achieve satisfactory results on the *eval* dataset in the final challenge evaluation.

[2] The *eval* results published in this paper have been computed using V2 dataset updated by the organisers of the challenge in May 2017.

[3] We evaluated Long Short-Term Memory (LSTM) networks with 1-3 recurrent layers and Convolutional Neural Networks (CNN) with 1-6 convolutional layers, followed by a softmax layer and cross-entropy training criterion. The frameworks used were TensorFlow [31] and Keras [32].

Table 1: *Equal error rates (EER's) for features extracted from different subbands from development set.*

Frequency range [Hz]	EER [%]				
	CQCC	Cepstrum	IMFCC	MFCC	LPCCres
16 – 8000	11.86	8.52	4.48	16.98	10.98
16 – 1000	42.56	38.48	35.69	27.48	20.60
1000 – 2000	47.03	38.98	42.19	41.05	9.60
2000 – 4000	42.26	39.60	36.99	38.30	9.77
4000 – 8000	7.23	5.27	**3.16**	**16.18**	**6.22**
6000 – 8000	**5.13**	**3.38**	4.16	16.76	6.37

Experiments with linear score fusion (using Bosaris toolkit [33]) for multiple classifiers improved overall EER on training and development data. However, different partitioning of development set for multiple-fold fusion training induced high variation in weights and resulting performance. Since high overfitting and sensitivity to the chosen dataset for training was observed, we decided not to apply such score fusion.

4. Results

Table 1 presents the EER results obtained for different features in a variety of frequency sub-bands. All features were modelled with the same GMM classifier described in Section 2.2. The 16 Hz limit resulted from dividing the sample rate by 2^{10}.

For CQCC and cepstrum features, we also performed the analysis separately for each octave below 1 kHz, but none of the sub-band results have reached less than 33% in terms of EER. The difference between our result for the full-band CQCC (the baseline), which amounted to EER=11.68%, and the result reported by the organisers, namely EER=10.75%, is a consequence of using a different classifier implementation.

In the full-band analysis, the best result of EER=4.48% was achieved for the IMFCC, which is a feature that emphasizes high frequencies. Compared to the baseline CQCC, it reduces EER by 63%. Secondly, all features from Table 1 exhibit a significant improvement in terms of the EER for $4-8$ kHz subband over the remaining sub-bands. The results for the high frequency analysis of different features clearly outperform the respective results for the full-band analysis. We conclude that the spoofing analysed in the challenge should be detected more effectively by the high-frequency countermeasures.

Let us discuss the EER results for the proposed LPCCres - a new feature obtained by combining the full-band LPCC based on the 35 LP filter coefficients with the sub-band cepstrum of the residual signal. As can be seen, the results are consistent and highly promising across different frequency bands, which is a consequence of using broadband LP coefficients. Using the LPCC, we were able to achieve the full-band EER=6.31%, which is the 2nd result compared to other broadband features. Furthermore, we tested LPC orders of 25, 30, and 40 and obtained the following EER results: 7.78%, 7.22% and 7.18%; hence significant improvements were not confirmed. The concatenation of full-band LPCC with the proposed LPCCres feature showed slight decrease of EER for the $4-8$ kHz band.

The outcome of the challenge evaluation is presented in Table 2, where we compare the results obtained for the development and evaluation datasets. As can be observed, the results obtained in the evaluation are significantly worse than the ones achieved on the development set.

In the following, we would like to briefly discuss whether a spoof detection system based on the discussed feature set is able to generalise to unseen data. We observe a 30% relative reduction of the EER with regard to the baseline system just by fine-tuning the input features. However, the difference between the performance on the *dev* set (5.13% EER) and the *eval* set (17.31% EER) is still substantial. It may be concluded that most likely only a subset of the spoofed recordings was significantly affected in the high-frequency sub-band. The reason behind this may be that the assumed high-frequency artefacts actually are not that severe in current devices. In addition, we believe that the limited number of spoofing conditions in the development set may have led to a strong over-fitting of the trained models, and consequently led to the overall poor generalisation in the evaluation.

Table 2: *Comparison of the results obtained on development (dev) and evaluation (eval) sets.*

Features	EER[%]	
	dev	eval
CQCC (full-band)	11.86	24.57
CQCC (6-8 kHz)	5.13	**17.31**
Cepstrum (6-8 kHz)	**3.38**	22.24
LPCCres (6-8kHz)	6.37	27.61

Future work will focus on more detailed examination of the proposed LPCCres features and investigation of their potential using the published evaluation dataset. To this end, a new optimized spoofing-detection filter-bank design is required, optimized sub-band LPC along with the presented sub-band LPCCres features should be examined, and an a-posteriori optimization of the most discriminative frequency analysis should be performed.

5. Conclusions

We investigated spectral alterations introduced in the process of replay spoofing and provided evidence that significant spoofing cues related to a multiple anti-aliasing filtering can be found at high frequencies. Several methods of high-frequency fine-grained parametrisation were scrutinised. The fine-tuned CQCC showed the strongest generalisation to unseen data, reducing the EER by 30%. The proposed approach does not solve the spoof detection problem completely, but it introduces a significant improvement over the baseline CQCC-GMM system.

6. Acknowledgements

This work was supported by the AGH University of Science and Technology Dean's Grant 15.11.230.336, by the National Science Centre under grant number DEC-2014/12/S/ST7/00265, and by means of statutory activity 11.11.230.017. We would also like to thank Mr Sahidullah and the organisers of ASVSpoof 2017 for performing additional evaluations.

7. References

[1] S. O. Sadjadi, S. Ganapathy, and J. Pelecanos, "The ibm 2016 speaker recognition system," in *Odyssey 2016*, 2016, pp. 174–180. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2016-25

[2] T. Hasan, G. Liu, S. O. Sadjadi, N. Shokouhi, H. Boril, A. Ziaei, A. Misra, K. Godin, and J. Hansen, "Utd-crss systems for 2012 nist speaker recognition evaluation," in *Proc. NIST SRE Workshop*, 2012.

[3] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Z. Ma, S. Cumani, O. Glembek, H. Hermansky, S. H. R. Mallidi, N. Mesgarani *et al.*, "Developing a speaker identification system for the darpa rats project." in *ICASSP*, 2013, pp. 6768–6772.

[4] 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.

[5] 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," 2016. [Online]. Available: http://www.spoofingchallenge.org/

[6] 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.

[7] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the*. IEEE, 2014, pp. 1–6.

[8] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–5.

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

[10] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1678–1681.

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

[12] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4. IEEE, 2011, pp. 1708–1713.

[13] J. Eargle, *Loudspeaker Handbook*. Springer, 2003. [Online]. Available: https://books.google.pl/books?id=Twu0oHE1ukgC

[14] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[15] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[16] 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.

[17] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[18] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

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

[20] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[21] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of crim system for the automatic speaker verification spoofing and countermeasures challenge 2015." in *INTERSPEECH*, 2015, pp. 2072–2076.

[22] A. Janicki, "Increasing anti-spoofing protection in speaker verification using linear prediction," *Multimedia Tools and Applications*, pp. 1–16, 2016.

[23] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. A. L. Thomsen, A. K. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings, 2017.

[24] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[25] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

[26] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[27] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," *Computer Speech & Language*, vol. 28, no. 1, pp. 121–140, 2014.

[28] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[29] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "The IBM 2016 speaker recognition system," *arXiv preprint arXiv:1602.07291*, 2016.

[30] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[32] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[33] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.