Interspeech 2018
2-6 September 2018, Hyderabad

Modulation Dynamic Features for the Detection of Replay Attacks

Gajan Suthokumar[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, v.sethu@unsw.edu.au, c.wijenayake@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

The development of automatic systems that can detect replayed speech has emerged as a significant research challenge for securing voice biometric systems and is the focus of this paper. Specifically, this paper proposes two novel features to capture the static and dynamic characteristics of the signal from the modulation spectrum, which complement short term spectral features for use in replay detection. The modulation spectral centroid frequency feature is proposed as a vector representation of the first order spectral moments of the modulation spectrum. In conjunction to this, the long term spectral average serves to capture the static characteristics of the modulation spectrum. The proposed system, employing a GMM back-end, was evaluated on the ASVSpoof 2017 dataset and found to yield an EER of 6.54%.

Index Terms: speaker verification, spoofing detection, score fusion, spectro temporal features, modulation spectrum

1. Introduction

Automatic speaker verification (ASV) has undergone rapid improvements in the recent decade but continues to show high vulnerability to spoofing attacks. Spoofing methods are categorized as speech synthesis (SS), voice conversion (VC), impersonation and replay [1]. Among these, replay attacks are arguably the most common ASV spoofing technique as they do not require the attackers to have any specialised knowledge, and can be mounted with relative ease using consumer devices. Countermeasures against replay attacks generally aim to exploit either the fact that replayed speech would be an exact reproduction of a previous speech utterance, differences in the speech transmission channel, or differences in the spectral properties of replayed speech. Playback audio detectors that successfully detect spoofing by comparing a new recording with previous attempts are described in [2]. Further approaches involve distinguishing between the transmission channels of genuine and replayed speech such as the detection of pop-noise [3], acoustic channel artefacts [4], and detection of far-field recording [5]. The short term features derived from linear filter banks (rectangular filter cepstral coefficients, spectral centroid magnitude coefficients [6], and instantaneous frequency features [7]), or from linear scaled spectrograms (single frequency features [8] and light CNN features [9]) have been shown to be superior to frequency warped features (constant-Q cepstral coefficients [10], inverse Mel frequency cepstral coefficients [11, 12]). In addition, different variants of neural network (NN) based systems [9, 13, 14] have also been investigated. Further appending the velocity (delta) and acceleration (delta-delta) features, which capture short-term dynamics between the nearest frames, have been shown to be beneficial. The fusion of different types of features and systems can also improve the results [6, 9, 15].

Moreover, long-term spectral statistics have also been shown to be helpful in replay attack detection [16]. However, the long-term spectro temporal dynamics is affected by noise and reverberation [17, 18, 19, 20] and has not been fully explored in the context of replay attack detection. Thus, in this paper, we mainly focus on long-term temporal dynamic information obtained from log spectrograms in terms of a modulation spectrum [18]. The motivations behind this work are: (1) the artefacts in long-term dynamics are not well captured by the short-term features, hence incorporation of a modulation spectrum's static and dynamic information would be beneficial; and (2) the modulation spectrum has been shown to be affected by reverberation and noisy conditions such as ambient noise and convolutional noise, which could be useful for replayed speech detection. Our final proposed system also incorporates a short term spectral feature to complement the modulation spectrum based features.

2. Features

2.1. Joint Acoustic Modulation Spectrum based features

2.1.1. The Motivation

The modulation spectrum characterizes the dynamics of the spectral content of the speech signal over a long duration. The modulation spectrum is also known to be a good indicator of speech intelligibility [19, 21], voice quality [18] and channel variability [17, 22]. We expect that replayed signals would include noise and reverberation, leading to a flatter modulation spectrum. The use of the modulation spectrum in synthesized speech detection has been previously investigated and shown to be promising [23].

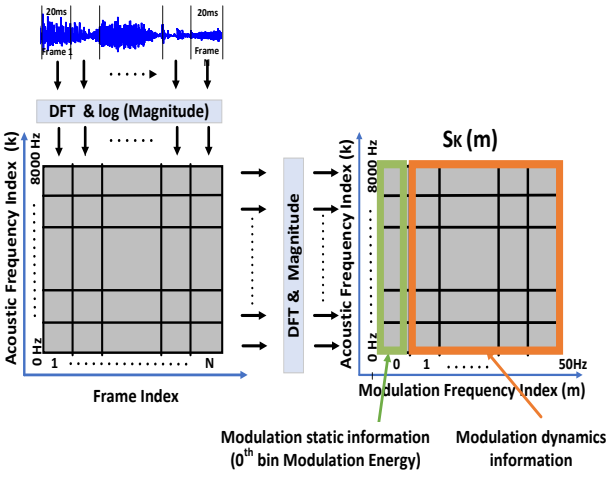

Figure 1: *Computation of short term log spectrogram (left), and joint acoustic modulation spectrum (right). Regions of representing static (green) and dynamics characteristics (orange) are indicated.*

10.21437/Interspeech.2018-1846

2.1.2. *Joint Acoustic Modulation Spectrum*

While a number of variations of the modulation spectrum have been proposed over the years [17, 20, 24, 25, 26], in this work we employ the 'Joint Acoustic Modulation Spectrum', as illustrated in Figure 1 (similar to that employed in [26, 27]). Specifically, it is estimated by taking the magnitude of the Discrete Fourier Transform (DFT) along time in a log magnitude spectrogram.

In the literature, two types of modulation spectra have been used: (a) the segment wise modulation spectrum, which is computed over a fixed number of frames; and (b) an utterance level modulation spectrum, which is calculated over all the frames in the speech utterance [28]. We use the latter approach since we are interested in the statistical effects present due to replay (transmission channels) that might present throughout the utterance. The utterance level modulation spectrum also leads to features that are more compact and independent of the length of the utterance.

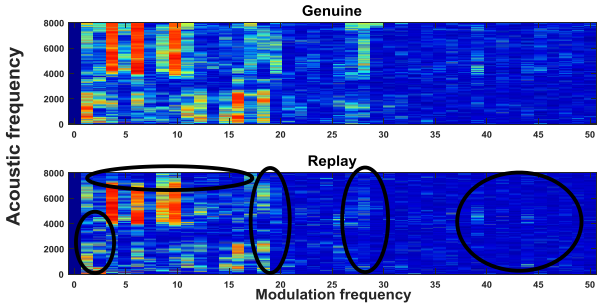

Figure 2: *Comparison of normalized modulation spectrum for genuine and replayed speech utterances of, "Birthday parties have cupcakes and ice cream". The 0th modulation bin is suppressed to highlight dynamic variations elsewhere.*

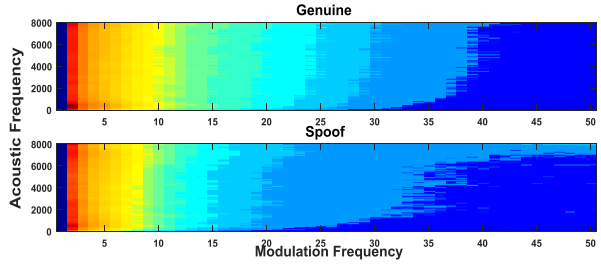

Figure 3: *Average Modulation Spectrum of genuine and spoofing utterances computed using the training set of ASVSpoof 2017 dataset. The 0th modulation bin is suppressed to highlight dynamic variations elsewhere.*

2.1.3. *Parameter Selection and Comparison*

The modulation spectrum was computed by initially estimating a spectrogram using 20ms hamming windows with 50% overlap after pre-emphasis of the speech signal with a 1024 point DFT. A second DFT is employed to obtain the modulation frequencies, employed as many points as there were frames in the utterance. As a result of the initial windowing, the highest modulation frequency present was 50Hz.

The modulation spectrum energy was normalized to maintain norms across utterances of different lengths. Figure 2 compares the joint acoustic modulation spectrum of a genuine speech utterance to that of a replayed version of that utterance. Replayed speech has variations in modulation components due to channel variations present in replayed speech. In addition, Figure 3 compares the average modulation spectrum estimated across all genuine speech in the training set to the average modulation spectrum across all spoofed speech in the training set. It is evident from this figure that there are differences in the total distribution of energy in the modulation spectrum between genuine and replayed speech, perhaps due to additional transmission channels present in replayed speech. Further differences in high acoustic frequency regions (i.e. the modulation energies) may be due to non-linearities at high frequencies in low quality playback and recording devices [11]. We propose two novel features to capture the static and dynamic characteristics of the signal from the modulation spectrum.

2.1.4. *Proposed Modulation Spectrum Based Features*

As depicted in Figure 4, we propose two novel features which are derived from the modulation spectrum (refer Figure 1), referred to as MCF-CC and MSE-CC. The modulation centroid frequency (MCF) computes the spectral centroid within each acoustic frequency bin to capture the modulation energy variation within. The spectral centroid (the first moment in the spectral domain) is a statistical measure which calculates the mean frequency [29]. This feature compactly represents the spread of energy across the modulation frequencies and is expected to capture the variation of modulation peak energy within acoustic frequency bins. The modulation centroid frequency (MCF) of the k^{th} acoustic frequency is

$$\mathrm{MCF}_k = \frac{\sum_{m=1}^{50} S_k(m) \cdot m}{\sum_{m=1}^{50} S_k(m)} \qquad (1)$$

where k is the acoustic frequency bin index, m is the modulation frequency bin index, and $S_k(m)$ is the normalized modulation energy (i.e. magnitude) corresponding to k^{th} acoustic frequency and m^{th} modulation frequency. The discrete cosine transform (DCT) is applied across the modulation centroid frequencies to create a compact utterance level feature, referred to as the MCF cosine coefficients (MCF-CC), as shown in Figure 4. The MCF-CC feature represents the utterance level dynamic information contained in the modulation spectrum.

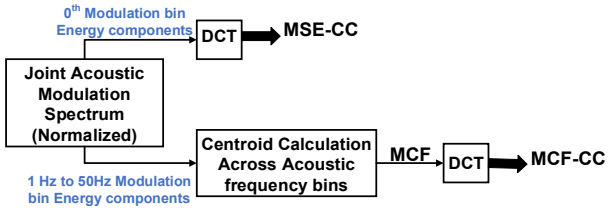

Figure 4: *MCF-CC and MSE-CC feature extraction from a modulation spectrum.*

As illustrated in Figure 4, the 0th modulation bin energies ($m = 0$) of the normalized modulation spectrum along the acoustic frequencies are retained as a feature vector, which we refer as named as modulation static energy (MSE), as

$$\mathrm{MSE}_k = S_k(0) \quad (2)$$

where k is the acoustic frequency bin index. Conceptually MSE carries the DC component information of the temporal trajectories of the log spectrogram. The DCT was performed to reduce the dimensionality of the MSE and gives rise to the compact MSE cepstral coefficient (MSE-CC) features. The MSE-CC features represent the utterance level static information from the modulation spectrum. It should be noted that mean normalization is not carried out at any stage.

2.2. Short Term Cepstral Coefficients (STCC)

In addition to the proposed modulation based features, we also estimate short term frame based features to complement the information extracted from the modulation spectrum. These features are extracted from the log spectrogram. Specifically, the first few DCT coefficients of the log magnitude spectrum from each frame are extracted as short-term features and herein referred to as short term cepstral coefficients (STCC).

3. Development and Experimental Setup

3.1. Dataset

The ASVSpoof 2017 version 1.0 dataset consisting of genuine recordings and their replayed versions as spoofed speech is used for our evaluations [30]. The dataset consist of three non-overlapping sets for training, development and evaluation. We used the development (dev) set to tune the system and the pooled set (train and dev) is used to train the genuine and spoof models for final evaluation.

3.2. Feature extraction

We carried out the experiments on 15, 30, 50 and 100 dimensions (the number of retained DCT coefficients) to determine the most informative dimensions by tuning on the development set. MCF-CC and MSE-CC features are chosen with 15 and 30 dimensions respectively. The STCC features are chosen to have 30 dimensions and are appended with their velocity and acceleration coefficients. Cepstral mean variance normalization (CMVN) is carried out for the STCC features.

3.3. Classifier

A Gaussian mixture model (GMM) back-end was employed for genuine and spoofed speech detection, which remains the de-facto technique [6, 9]. The GMMs were trained using the expectation maximization (EM) criterion to obtain the maximum likelihood estimate with random initialization. The number of Gaussian mixtures components for MCF-CC, MSE-CC and STCC were chosen as 4, 4 and 512 respectively, determined on the dev set.

3.4. Score and feature level fusion

The features associated with each method are expected to be complimentary to each other and score level fusion was performed using the BOSARIS toolkit [31]. Linear fusion parameters are learned from the development set. Feature level concatenated feature (15 dimension of MCF-CC and 30 dimension of MSE-CC) was also investigated for the two proposed modulation features using a 4 mixture GMM back-end as they are extracted from two non-overlapping regions from the modulation spectrum.

4. Results & Discussion

We present the results of our proposed systems on ASVSpoof 2017 version 1.0 corpus [30] to compare with baseline systems evaluated in this dataset. Henceforth S1, S2 and S3 refer to systems using MCF-CC, MSE-CC and STCC features, respectively (see Table 1). We also present our proposed system evaluation results on ASVSpoof 2017 version 2.0 corpus with latest protocols [10]. A subsequent ASVSpoof 2017 version 2.0 [10] was released in early 2018 to fix the data anomalies detected in the version 1.0 of the corpus. In addition, meta data which describes the playback device, recording device and the different acoustic environment condition used in the evaluation set is also released. All results are provided on ASVSpoof 2017 version 1.0 dataset except where noted.

Two dimensional t-SNE projections of both modulation features are shown in Figure 5. Both the features show discrimination ability between genuine and spoofed speech. The distribution of the evaluation scores for the individual (S1 and S2) and fused systems (S1+S2 at feature and score levels) are shown in Figure 6. Both features are highly complementary and benefit from fusion at both feature and score levels. Scatter plots of the evaluation scores, shown in Figure 7, were investigated to study the complimentary information present in system S1 with S2 and S3. The regions of interest are Regions I and II (R1 and R2), which correspond to an incorrect classification (false acceptance and miss rate). Systems S1 and S2, or S1 and S3 together are more robust and have high complimentary information (i.e. misclassification in S1 is complimented by S3 to make the correct decision and vice-versa.).

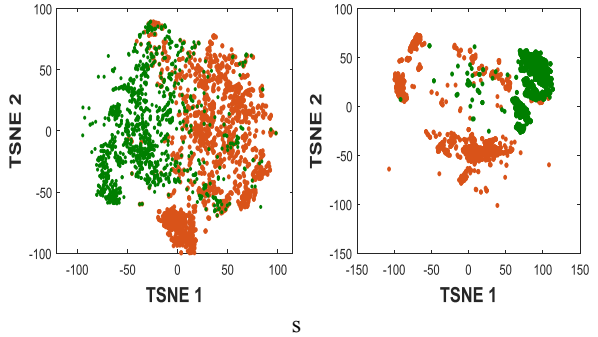

Figure 5: *t-SNE two dimensional feature space of MCF-CC (left) and MSE-CC(right) of genuine (orange) and spoofed (green) speech on the training set.*

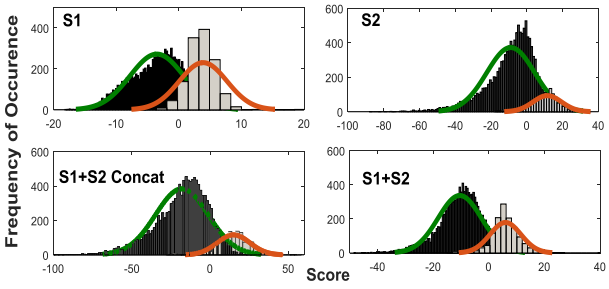

Figure 6: *Evaluation score histogram distribution for systems S1, S2, S1+S2 Concat (feature fusion), and S1+S2 (score fusion), for genuine (orange) and spoof (green) trials.*

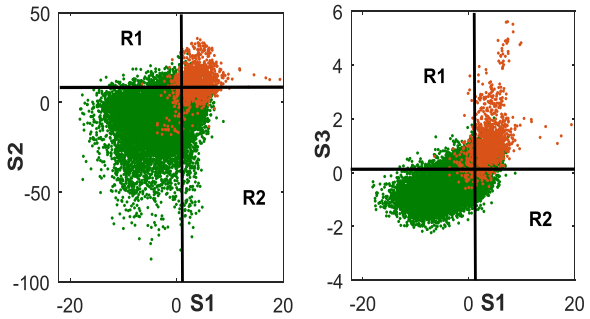

Figure 7: *Scatter plot of system evaluation scores for genuine (orange) and spoofed (green) utterances: systems S1 vs S2 (left) and S1 vs S3 (right). Vertical and horizontal lines indicate the decision threshold for the relevant system.*

Table 1 presents the performances of the individual systems on the evaluation set in terms of equal error rate (EER). Table 2 compares fused systems with the baseline systems. The proposed modulation features system (S1+S2) performs significantly well with an EER of 7.20% and 7.97% with score and feature level fusion respectively which depicts the features' complimentary nature. Our best proposed system (S1+S2+S3) yields an EER of 6.54%, outperforming all other systems. Our proposed system results improved further to an EER of 6.32% in version 2.0 dataset. To the best of the authors' knowledge, this is the best performance reported in this standard corpus. Table 3 presents the results of our best system (S1+S2+S3) for different replay conditions according to the threat level as defined in [10]. The proposed system outperforms previously reported results [10] under all nine threat conditions.

Table 1: *Individual system evaluation results in terms of % EER (All results are provided on ASVSpoof 2017 version 1.0 dataset).*

Technique	% EER
S1: MCF-CC	12.92
S2: MSE-CC	11.97
S3: STCC	11.27

Table 2: *Comparaison of evaluation results with baseline systems in terms of % EER (All results are provided on ASVSpoof 2017 version 1.0 dataset except where noted).*

	System Description	% EER
Baseline	Light CNN fused systems [9]	6.73
	Light CNN + GMM [9]	7.37
	CNN +RNN+ GMM [9]	10.69
	SCMC+GMM [6]	11.49
	RFCC+GMM [6]	11.90
Proposed	S1+S2 (feature fusion)	7.97
	S1+S2 (score fusion)	**7.20**
	S1+S3	9.21
	S2+S3	8.25
	S1+S2+S3	**6.54**
	S1+S2+S3 (version 2.0 dataset)	6.32

Table 3: *The best proposed system (S1+S2+S3) evaluation results (bold) and best reported results[10] (In parentheses) for low, medium and high threat conditions in terms of % EER, on version 2.0 dataset.*

Conditions	Low	Medium	High
Environment	**6.36** (16.68)	**5.97** (18.73)	**8.68** (21.86)
Playback	**5.12** (16.64)	**5.85** (16.44)	**7.76** (18.37)
Recording	**3.53** (10.80)	**6.68** (15.69)	**7.16** (17.77)

5. Conclusions

We have proposed and investigated two utterance level features extracted from the modulation spectrum that help to identify replayed speech based on the static and dynamic characteristics of the speech signal. The two features complement each other and are further complemented by short term cepstral features. A fused system was then implemented to combine all three features with score level fusion. This system is less complex compared to state-of-the-art systems, as it uses two utterance level features and a small number of Gaussian mixtures as a classifier. Experimental results for the modulation features system obtained on the standard ASVSpoof 2017 corpus show that they perform significantly better compared to the state-of-the-art systems, with an EER of 7.20%, and the system that fused the two modulation features with short term cepstral features further improved the EER to 6.54%.

6. References

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

[2] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," *2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014*, pp. 92–96, 2014.

[3] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice Liveness Detection for Speaker Verification Based on a Tandem Single / Double-Channel Pop Noise Detector," pp. 259–263, 2016.

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

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

[6] 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, pp. 7–11.

[7] 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*, 2017, pp. 12–16.

[8] K. N. R. K. R. Alluri, S. Achanta, and S. R. Kadiri, "Detection of Replay Attacks using Single Frequency Filtering Cepstral Coefficients," in *Interspeech*, 2017, pp. 2596–2600.

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

[10] M. Todisco, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey (Accepted)*, 2018.

[11] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," in *Interspeech*, 2017, pp. 27–31.

[12] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1195–1198.

[13] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection," in *Interspeech*, 2017, pp. 102–106.

[14] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay Attack Detection using DNN for Channel Discrimination," in *Interspeech*, 2017, pp. 97–101.

[15] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Independent Modelling of High and Low Energy Speech Frames for Spoofing Detection," in *Interspeech*, 2017, pp. 2606–2610.

[16] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2098–2111, 2017.

[17] T. H. Falk and W. Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, 2010.

[18] X. Xiao, E. S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1662–1674, 2008.

[19] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Commun.*, vol. 45, no. 2, pp. 101–113, 2005.

[20] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, 2011.

[21] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[22] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Natl. Acad. Sci.*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[23] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Icassp*, 2013, pp. 7234–7238.

[24] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neurosci. Biobehav. Rev.*, vol. 81, pp. 181–187, 2017.

[25] T. Kinnunen, K. A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," *Proc. Speak. Odyssey*, p. 30, 2008.

[26] T. Kinnunen, "Joint Acoustic-Modulation Frequency for Speaker Recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 1, p. 665–668s.

[27] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 4, no. 3, pp. 605–608, 2007.

[28] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 768–783, 2016.

[29] D. Cabrera, S. Ferguson, and R. Maria, "Using sonification for teaching acoustics and audio," in *Proceedings of the 1st Australasian Acoustical Societies' Conference (ACOUSTICS)*, 2006, no. November, pp. 383–390.

[30] T. Kinnunen *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Interspeech*, 2017, pp. 2–6.

[31] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," Apr. 2013.