AUDITORY INSPIRED SPATIAL DIFFERENTIATION FOR REPLAY SPOOFING ATTACK DETECTION

Buddhi Wickramasinghe[1,2], Eliathamby Ambikairajah[1,2], Julien Epps[1,2], Vidhyasaharan Sethu[1], Haizhou Li[3]

[1]School of Electrical Engineering and Telecommunications, UNSW, Australia
[2]ATP Research Laboratory, DATA61, CSIRO, Australia
[3]Dept of Electrical & Computer Engineering, NUS, Singapore

ABSTRACT

The security of Automatic Speaker Verification systems is greatly threatened by spoofing attacks of various kinds. Among them, replay attacks are noteworthy due to the ease with which they can be employed. Most countermeasures for replay attacks use subband features based on parallel filter banks. This paper explores the effect of 'spatial differentiation' used in auditory system modelling to improve frequency selectivity and hence provide a more selective front-end for replay attack detection. Experiments were done using a parallel filter bank consisting of simple 2nd order IIR bandpass filters following which, processing analogous to spatial differentiation was employed to obtain higher order stable IIR filters, in turn leading to highly selective filter banks. Two novel features based on spatially differentiated higher order filter bank have been proposed. Together they yield a relative improvement of 29.9% in replay speech detection over a constant Q transform based baseline system, when evaluated on the ASVspoof 2017 Version 2.0 database.

***Index Terms*— Automatic speaker verification, Anti-spoofing, ASVspoof 2017, Spatial differentiation**

1. INTRODUCTION

Automatic Speaker Verification (ASV) has gained much attention as a biometric authentication technique in the past few decades. However, with the increased implementation of ASV in practical contexts, different forms of attacks which try to deceive ASV systems are also becoming prominent. These attacks include identical twins [1], impersonation, voice conversion, speech synthesis and replay attacks [2]. Among them, replay attacks can be mounted rather easily using consumer devices without much technical expertise. A replay attack simply means an attacker recording the voice of a verified speaker and playing it back to the ASV system to gain access to secured content. The number of available recording and playback devices and possible recording environments is numerous. Therefore, effective countermeasures should be able to deal with varied unknown acoustic conditions. Hence, developing generalizable countermeasures for replay attack detection has become one of the major challenges faced by the ASV research community.

State-of-the-art research on replay attack detection are based on the ASVspoof 2017 database [3]. Many systems which explore various front-end features and back-end classifiers have been proposed based on this database. Most features to date for replay spoofing detection are based on the speech magnitude spectrum and employ subband decomposition using conventional filter banks such as triangular, rectangular or Gabor filters. Some examples include Mel Frequency Cepstral Coefficients (MFCC) and Rectangular Filter Cepstral Coefficients (RFCC) [4]. Other features include Amplitude Modulation (AM) [5], subband spectral centroid magnitude and subband spectral centroid frequency, the latter two of which contain some information on the subband energy distribution of a signal [4]. Mel and linear filter bank based spectral slopes have been used to extract low and high frequency information respectively showing improved detection accuracy over constant Q transform based features [6]. Investigation of frequency selectivity of filters in the context of replay attack detection has been limited to the number of filters and frequency scale used in filter banks.

Some countermeasures have taken rather different approaches, such as using Empirical Mode Decomposition (EMD), to decompose a signal into subbands [7]. Phase-based features such as modified group delay [8, 9] and frequency modulation (FM) [10] have also been used as front-end features. Speech production source information has also proven effective for replay attack detection [11, 12]. Long-term time frequency representations such as modulation spectra have also shown great promise [13]. Although many novel features have been proposed, the best performing systems for replay attack detection are fusions of many systems [14].

Feature extraction processes motivated by the human auditory system have been used in various fields such as speech recognition [15] and speaker recognition [16]. For example, computational auditory scene analysis based on Gammatone filters models human ability to selectively listen to channels in a multichannel situation [15]. In [16] a physiologically based auditory periphery model has been used to extract features for robust speaker identification. The importance of frequency selectivity in the inner ear for speaker identification has been investigated here. However, auditory system-based concepts are not widely used in replay attack detection even though techniques such as constant Q transform [17] which are analogous to auditory perception have shown consistently good performance [18]. Sailor et al. have used convolutional restricted Boltzman machine to learn an auditory-like filter bank to extract AM-FM features. But the auditory aspect of the filters has not been investigated in depth [19]. Since the auditory system has a remarkable ability to analyze and recognize complex sounds, adopting more auditory concepts in replay attack detection may improve performance.

The basilar membrane located within the cochlea of the human auditory system produces mechanical displacements as responses to input pressure variations [20]. The point with highest displacement is determined by the frequency of the input. 'Spatial

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

differentiation' has been used to model the mechanical coupling within the basilar membrane [21]. These frameworks have modelled the human cochlear using a transmission line system, where the basilar membrane is represented as a cascade of filters tuned to different frequencies [22]. Here, the word 'spatial' comes from the fact that the filtering effect is related to the mechanical displacement along the length of the basilar membrane. It has been established that spatial differentiation increases the frequency selectivity of filters in an auditory filter model [23]. Spatial differentiation has been used to model fluid coupling, in turn giving rise to additional sharpening mechanisms in the basilar membrane [21, 22].

This paper proposes using spatial differentiation proposed in auditory system modelling as a technique to improve the selectivity of filters of a parallel filter bank in feature extraction for replay attack detection.

2. SPATIAL DIFFERENTIATION

Define a filter bank of N 2nd order infinite impulse response (IIR) bandpass filters $\mathcal{F}^{(0)} = \left\{H_i^{(0)}(z); i = 1, \dots, N\right\}$, with each filter defined as:

$$H_i^{(0)}(z) = \frac{1-z^{-2}}{1-2r_i\cos\theta_i z^{-1}+r_i^2 z^{-2}} \tag{1}$$

where, i denotes the filter number, the (0) in the superscript denotes that this is the undifferentiated filter bank, and θ_i and r_i denote the pole angle and radius of the i^{th} filter respectively. This filter can be considered as a simple approximation of an auditory filter. A new filter bank can be obtained by spatial differentiation, $\mathbf{S}\{\cdot\}$, as follows:

$$\mathcal{F}^{(j+1)} = \mathbf{S}\{\mathcal{F}^{(j)}\} = \left\{H_i^{(j+1)}(z);\ i = 1, \dots, N-1\right\}$$

where,

$$H_i^{(j+1)}(z) = H_{i+1}^{(j)}(z) - H_i^{(j)}(z) \tag{2}$$

and j here denotes the order of spatial differentiation.

To show the effect of spatial differentiation, a parallel filter bank consisting of 2nd order IIR bandpass filters equally spaced in mel frequency scale were created and the magnitude response of one of the filters in the filter bank is shown before and after spatial differentiation. Specifically, in this example $N = 80$ and the magnitude response of the 55th filter (initial centre frequency of 3349.3Hz) is shown in Figure 1 along with the magnitude responses of the 55th filter after 3 and 6 spatial differentiations.

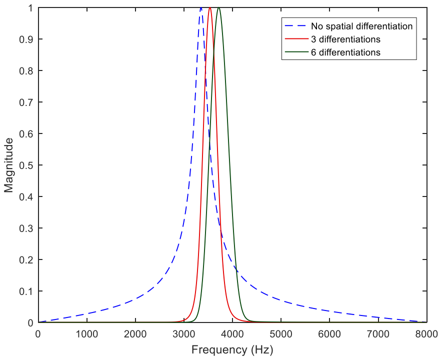

Figure 1: *Variation of frequency response with spatial differentiation for 55th filter of the filter bank. The response has become sharper with spatial differentiation.*

Note that the gains of the filters were set to 1 for easy comparison. Some observations can be made using the above figure: The centre frequency of each filter has increased with each differentiation, and the filter shape becomes sharper with steeper slope.

It can be seen that the transfer function of the filters in the filter bank after the first spatial differentiation is increased to 4th order and is given by:

$$H_i^{(1)}(z) = \frac{k_i z^{-1}(1-z^{-2})}{1+c_{i_1}z^{-1}+c_{i_2}z^{-2}+c_{i_3}z^{-3}+c_{i_4}z^{-4}} \tag{3}$$

where,

$$k_i = -4r\left(\frac{\theta_{i+1}-\theta_i}{2}\right) sin\left(\frac{\theta_i+\theta_{i+1}}{2}\right) \tag{4}$$

and $c_{i_1} = -4rcos\left(\frac{\theta_{\mathrm{i}+1}+\theta_{\mathrm{i}}}{2}\right)$, $c_{i_2} = 2r^2(1 + 2\cos(\theta_{\mathrm{i}+1})\cos(\theta_{\mathrm{i}}))$
$c_{i_3} = -4r^3cos\left(\frac{\theta_{\mathrm{i}+1}+\theta_{\mathrm{i}}}{2}\right)$, $c_{i_4} = r^4$ (5)

Several assumptions were made when obtaining this expression. Namely, since the two filters are adjacent in frequency, their pole radii are almost equal. Hence, $r_i \approx r_{i+1}$. Since the filters are close to each other in frequency, $\left(\frac{\theta_{i+1}-\theta_i}{2}\right)$ is assumed small enough to satisfy $sin\left(\frac{\theta_{i+1}-\theta_i}{2}\right) \approx \left(\frac{\theta_{i+1}-\theta_i}{2}\right)$ and $cos\left(\frac{\theta_{i+1}-\theta_i}{2}\right) \approx 1$.

Magnitude responses obtained after spatial differentiation showed that the centre frequency of the new filter after spatial differentiation is equal to the mean of centre frequencies of the two adjacent filters prior to spatial differentiation. This shift can be observed in Figure 1 as well. The increased sharpness of frequency responses after spatial differentiation can be explained by the increased filter order. Hence, the selectivity of the filter can be considered to have increased as well. Following the above approach, further spatial differentiations will give bandpass filters of order 8, 16, 32, 64, 128 and so on. Although the filter order has increased, the filters remain stable. Consequently, spatial differentiation over a simple parallel 2nd order bandpass filter bank can be used to obtain stable higher order IIR filters.

Finally, it should be noted that rather than directly implementing these higher-order filter banks for feature extraction, only the 2nd order filter bank was implemented, and the output signals of these filters were 'spatially differentiated' iteratively (refer Figure 2). This allows instabilities due to implementation of high order IIR filters to be avoided and is directly inspired by auditory modelling [24].

3. FEATURE EXTRACTION

We propose two novel features which can be used to extract discriminating information from the filtered and spatially differentiated speech signals. Features extracted from the smoothed envelope of a signal have given promising results for replay attack detection [25]. Hence, slowly varying spectral envelopes of the subband signals were used to extract features in this paper. The feature extraction process is shown in Figure 2. Speech signals were first filtered using a filter bank consisting of N 2nd order bandpass filters. Next, k spatial differentiations were applied iteratively on each filtered signal. The absolute value of each subband signal was calculated (i.e. full wave rectification) and each absolute signal was windowed into a set of overlapping frames. This signal was converted to the frequency domain by applying the DFT to each frame. Since the resulting signal contains frequency components of the envelope, it is of low frequency. Next, following two features were extracted from the output.

3.1 Spectral envelope centroid frequency (CF)

The centroid frequency (*CF)* of the spectral envelope can be defined as the weighted average frequency of the spectral envelope of the selected frame, where the weights are the magnitude of each frequency component. *CF* is representative of the distribution of energy in each envelope:

$$CF_k = \frac{\sum_{f=f_l}^{f_u} f.|W[f]|}{\sum_{f=f_l}^{f_u} |W[f]|} \qquad (6)$$

where f is frequency of each component, f_l and f_u are the lower and upper frequency limits of the subband signal and $|W[f]|$ is the spectral envelope of the subband signal.

3.2 Spectral envelope centroid magnitude (CM)

The other feature proposed is the centroid magnitude (*CM*) of the spectral envelope. This feature can be defined as the weighted average magnitude of the envelope under consideration. The weights are frequencies of each magnitude component. *CM* is the magnitude at the frequency position given by *CF*. Hence, the two features may contain complementary information. This feature can be considered as the frequency domain counterpart of temporal centroid amplitude feature proposed in our previous work [25].

$$CM_k = \frac{\sum_{f=f_l}^{f_u} f.|W[f]|}{\sum_{f=f_l}^{f_u} f} \qquad (7)$$

4. EXPERIMENTAL SETUP

4.1 Database

Experiments were conducted on the ASVspoof 2017 Version 2.0 database [18]. This database is a modified version of the ASVspoof 2017 database [3] which consists of genuine speech utterances and spoofed speech utterances which were created by replaying and recording genuine utterances using various playback and recording devices in varied acoustic conditions. All signals were sampled at 16 kHz. The evaluation set of the database includes many utterances generated under replay conditions which are unseen in the training and development sets.

4.2 Front-end configuration

Parameters of the system were tuned using preliminary experiments on the development set of the database. Accordingly, a mel frequency scaled filter bank consisting of 80 2nd order IIR bandpass filters previously designed (equation (1)) was used to filter the signals. Pole radius and pole angle of each filter were obtained by first calculating centre frequencies and bandwidths of the filters. It is clear that each spatial differentiation reduces the number of subband signals by one. To maintain the dimension at 80, a zero padding was introduced for the highest frequency subband after each intermediate differentiation. After the final differentiation, the highest frequency band was padded with the values of the subband before it. After spatial differentiation, the absolute value of each signal was Hamming windowed into 20 ms long frames with a 10 ms overlap. A DFT with 320 samples (same as frame length) was applied on each frame. Since the spectral envelopes contain low frequency content, *CF* and *CM* values were extracted for each frame per subband using frequency components only up to 950 Hz. Delta and acceleration coefficients of each feature were concatenated with them to provide additional dynamic information.

The discrete cosine transform (DCT) was applied across subbands to the log of the *CM* feature set to reduce the correlation between values. Based on development set results, only the first 40 coefficients of the feature were used for classification. Since *CF* is a frequency domain feature, log and DCT were not applied to these features. Hence, the dimension of *CF* feature was kept as 80. Utterance-wise mean-variance normalization was applied across frames to align genuine and replayed speech distributions to a common scale.

4.3 Back-end classifier

Based on feature dimensions and database size, a 2-class Gaussian Mixture Model (GMM) classifier with 512 mixtures was used as the back-end classifier for the experiments. Two GMMs, one modelling genuine data and one modelling spoofed data were

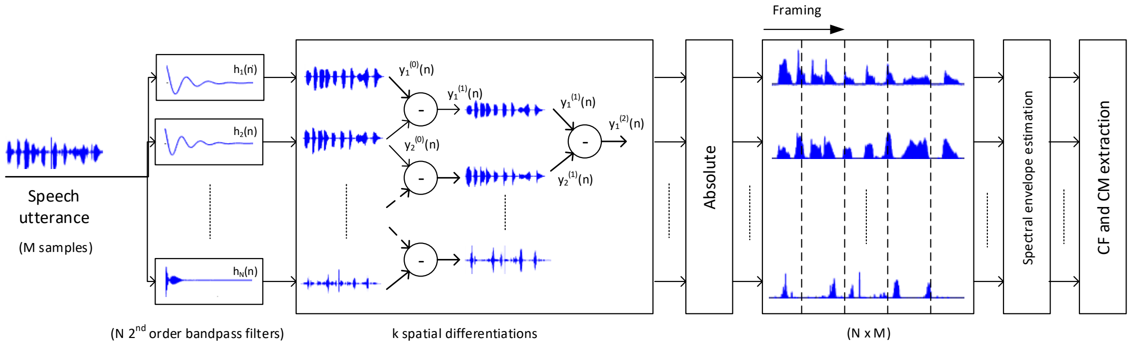

Figure 2: *CF and CM feature extraction process: k spatial differentiations were applied to speech filtered using N 2nd order bandpass IIR filters. Features were extracted from the spectral envelope estimated from the full-wave rectfied signal;* $\left\{y_i^{(0)}(n); i = 1, \ldots, N-1\right\}$ *were the non spatial differentiated filtered outputs, where (0) denotes the order of differentiation.*

Table 1: *EER (%) using the evaluation set of ASVspoof 2017 Version 2.0 database using two proposed features for different numbers of spatial differentiations*

Feature	EER (%)			
	No spatial differentiation	**2 spatial differentiations**	**4 spatial differentiations**	**6 spatial differentiations**
CF + Δ + ΔΔ	20.67	14.31	11.92	10.84
CM + Δ + ΔΔ	20.24	15.22	11.61	10.93
CM + Δ + ΔΔ (Tri)	13.57	12.42	12.03	13.29

created using the MSR Toolkit [26]. A log-likelihood ratio was calculated for each test utterance as the classification score.

5. EXPERIMENTAL RESULTS

Utterances from training and development sets were pooled together when training the GMMs for experiments on the evaluation set. Equal error rate (EER) was used as the performance evaluation metric in all experiments. First, the detection performances of each individual feature along with their delta and acceleration coefficients were evaluated with and without spatial differentiation.

Classification accuracies for 2, 4 and 6 spatial differentiations were obtained. Results are given in Table 1, from which it can be seen that EER consistently reduces with an increased number of spatial differentiations for both features. The reason for such improved performance could be the increased selectivity of filters. Similar performance gains for both features show the consistency of the technique.

Same experiments with *CM* feature were carried out using a triangular filter bank with same centre frequencies as the 2nd order bandpass filter (denoted 2nd order BPF) for comparison. Speech signals were converted to the frequency domain by applying DCT, and filtering was done in a frame-wise manner. Since spatial differentiation was applied to time domain signals, inverse DCT was applied prior to spatial differentiation. DCT was chosen over DFT because DCT provides a real-valued frequency domain signal. *CM* features were extracted similarly as described in Section 3. Results are given in Table 1 (denoted CM + Δ + ΔΔ (Tri)). With no spatial differentiation, triangular filter system performed better than the 2nd order BPF system. Although spatial differentiation slightly reduced the error rate up to 4 differentiations, the effect was not as significant as in the 2nd order BPF systems. Importantly, EER increased after applying 6 differentiations. Hence, it is seen that the improvement that can be achieved using triangular filters is limited compared to 2nd order bandpass filters.

Next, the *CF* and *CM* features along with their delta and acceleration coefficients obtained after 6 spatial differentiations (denoted 6 SD) using 2nd order BPF were concatenated to form a new feature and the replay detection performance of this system was also assessed. The two best performing individual systems in Table 1 were fused at the score level using the FoCal Toolkit [27] to evaluate the complementary nature of the features further, with results shown in Table 2.

Experimental results were compared with three baseline systems. The first baseline system (B1) used here [18] comprised 19th order Constant Q Cepstral Coefficients (CQCC), along with their delta and acceleration coefficients with log energy coefficients of the signals also appended as the front-end feature. The second baseline system (B2) was based on temporal centroid amplitude (TC) feature proposed in our previous work [25]. This feature was extracted from the temporal envelope estimated using frequency domain linear prediction. The final baseline (B3) was a score-level fusion of three systems based on features extracted from the modulation spectrum of speech a signal [13]. The features used in these three systems were modulation centroid frequency cosine coefficients (MCF-CC), modulation static energy cosine coefficients (MSE-CC) and short-term cepstral coefficients (STCC). All three systems used GMMs as their back-ends.

Table 2: *EER (%) values before and after fusion*

	System	EER (%)
B1	CQCC [18]	12.24
B2	TC [25]	14.89
B3	MCF-CC+MSE-CC+STCC [13]	6.54
S1	CF + Δ + ΔΔ (6 SD)	10.84
S2	CM + Δ + ΔΔ (6 SD)	10.93
	S1 + S2 (Feature level)	**8.58**
	S1 + S2 (Score level)	8.80

From Table 2, as single systems, S1 and S2 both performed better than baseline systems B1 and B2. Both fusion approaches show almost similar performance. Fusion of the two systems has brought the error rate down further, showing the complementary nature of the two features. However, the fused systems showed higher EER than B3, which may be because B3 uses long-term as well as short-term modulation information from speech signals. Hence, the increased filter order due to spatial differentiation may have led to high selectivity of filters, resulting in improved replay attack detection performance relative to selected baselines.

6. CONCLUSION

This paper investigates the effect of using spatial differentiation on filter outputs of a parallel filter bank as a means to improve replay spoofing attack detection accuracy. It was shown that spatial differentiation can be used to obtain stable higher order IIR filters, starting with a simple 2nd order IIR bandpass filter. Hence, this method can be used to easily develop filters with more selectivity. The improved performance of the filters was evaluated by applying spatial differentiation as a pre-processing technique in a front-end feature extraction method for replay attack detection. Two simple systems with conventional GMM back-ends which outperformed some baselines were developed. Fusion of the systems brought the error rate down further. Since spatial differentiation improves filter selectivity, it could be adopted in many feature extraction processes without limiting to replay attack detection.

REFERENCES

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE,* vol. 64, no. 4, pp. 475-487, 1976.

[2] 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.

[3] T. Kinnunen *et al.*, "ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training,* vol. 10, no. 1508, p. 1508, 2017.

[4] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–Results on the ASVspoof 2017 Challenge," *Proc. Interspeech 2017,* pp. 7-11, 2017.

[5] M. Kamble and H. Patil, "Novel Variable Length Energy Separation Algorithm Using Instantaneous Amplitude Features for Replay Detection," *Proc. Interspeech 2018,* pp. 646-650, 2018.

[6] M. Saranya and H. A. Murthy, "Decision-level feature switching as a paradigm for replay attack detection," *Proc. Interspeech 2018*, pp. 686-690, 2018.

[7] P. Tapkir and H. Patil, "Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection," *Proc. Interspeech 2018,* pp. 721-725, 2018.

[8] F. Tom, M. Jain, and P. Dey, "End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention," *Proc. Interspeech 2018,* pp. 681-685, 2018.

[9] D. Li *et al.*, "Multiple Phase Information Combination for Replay Attacks Detection," *Proc. Interspeech 2018,* pp. 656-660, 2018.

[10] T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, and J. Epps, "Detection of Replay-Spoofing Attacks Using Frequency Modulation Features," *Proc. Interspeech 2018,* pp. 636-640, 2018.

[11] S. Jelil, R. K. Das, S. M. Prasanna, and R. Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," *Proc. Interspeech 2017,* pp. 22-26, 2017.

[12] S. Jelil, S. Kalita, S. M. Prasanna, and R. Sinha, "Exploration of Compressed ILPR Features for Replay Attack Detection," *Development,* vol. 8, no. 760, p. 950, 2018.

[13] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation Dynamic Features for the Detection of Replay Attacks," *Proc. Interspeech 2018,* pp. 691-695, 2018.

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

[15] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4625-4628: IEEE.

[16] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PloS one,* vol. 11, no. 7, p. e0158520, 2016.

[17] 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.

[18] H. Delgado *et al.*, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 296-303.

[19] H. Sailor, M. Kamble, and H. Patil, "Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection," *Proc. Interspeech 2018,* pp. 666-670, 2018.

[20] S. Seneff, "Pitch and spectral analysis of speech based on an auditory synchrony model," *Ph.D. Thesis of Speech Communication Group, MIT,MA*, 1985.

[21] S. A. Shamma and K. A. Morrish, "Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea," *The Journal of the Acoustical Society of America,* vol. 81, no. 5, pp. 1486-1498, 1987.

[22] E. Ambikairajah, N. D. Black, and R. Linggard, "Digital filter simulation of the basilar membrane," *Computer Speech and Language,* vol. 3, pp. 105-118, 1989.

[23] J. Hall, "Spatial differentiation as an auditory''second filter'': Assessment on a nonlinear model of the basilar membrane," *The Journal of the Acoustical Society of America,* vol. 61, no. 2, pp. 520-524, 1977.

[24] M. C. Lang, "Least-squares design of IIR filters with prescribed magnitude and phase responses and a pole radius constraint," *IEEE Transactions on Signal Processing,* vol. 48, no. 11, pp. 3109-3121, 2000.

[25] B. Wickramasinghe, S. Irtza, E. Ambikairajah, and J. Epps, "Frequency Domain Linear Prediction Features for Replay Spoofing Attack Detection," *Proc. Interspeech 2018,* pp. 661-665, 2018.

[26] 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 1.4*, pp. 1-32, 2013.

[27] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores—Tutorial and user manual—," *Software available at http://sites. google. com/site/nikobrummer/focalmulticlass,* 2007.