INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Biologically Inspired Adaptive-Q Filterbanks for Replay Spoofing Attack Detection

Buddhi Wickramasinghe[1,2], *Eliathamby Ambikairajah*[1,2], *Julien Epps*[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW, Australia
[2] ATP Research Laboratory, DATA61, CSIRO, Australia
b.wickramasinghe@unsw.edu.au

Abstract

Development of generalizable countermeasures for replay spoofing attacks on Automatic Speaker Verification (ASV) systems is still an open problem. Many countermeasures to date utilize bandpass filters to extract a variety of frequency band-based features. This paper proposes the use of adaptive bandpass filters, a concept adopted from human cochlear modelling to improve detection performance. Gains of filters used for subband based feature extraction are adaptively adjusted by varying their Q factors (Quality factor) as a function of input signal level to boost low amplitude signal components and improve the front-end's sensitivity to them. This method is used to enhance information embedded in speech signals such as device channel effects which could be instrumental in distinguishing genuine speech signals from replayed ones. Three features extracted using the adaptive filter process yielded performance improvements over other auditory concepts-based baselines, showing the potential of using an adaptive filter mechanism for replay spoofing attack detection.

Index Terms: ASVspoof 2017, Replay attacks, Anti-spoofing, adaptive filtering

1. Introduction

Automatic Speaker Verification (ASV), a form of biometric authentication, uses a person's speech signal to verify his/her identity. Although ASV has gained popularity in many areas, its vulnerability to spoofing has been acknowledged [1]. Spoofing attacks on ASV systems are categorized into impersonation, speech synthesis, voice conversion, replay attacks [1] and identical twins [2]. Among them, replay attacks stand out due their simple nature. An attacker mounting a replay attack records the target speaker's speech and plays it back to an ASV system to gain access maliciously. The characteristics of replayed speech vary according to the devices used to record and playback target speech signals [3]. Consequently, effective countermeasures should be able to robustly identify varied characteristics of replay attacks.

Earlier research has used similarity scores to decide if an input is an exact reproduction of an earlier access attempt [4, 5]. Other countermeasures include discriminating replayed speech from genuine speech using the channel characteristics of replayed speech and added reverberation [6, 7]. However, the reliability of these early countermeasures is questionable due to the limited amount of data used in the experiments. With the release of ASVspoof 2017 database [8], many other countermeasures have been proposed. Front-end features include common cepstrum based features such as Mel Frequency Cepstral Coefficients (MFCC), Spectral Centroid Magnitude (SCM) features and Rectangular Filter Cepstral Coefficients (RFCC) [9]. Apart from these, other novel magnitude spectrum-based features have also been proposed. These include inverse MFCC features [10], Mel-filterbank slope (MFS) and Linear-filterbank slope (LFS) [11]. Most of the above-mentioned features extract discriminating information from the magnitude spectrum of a speech signal.

Speech phase-based features and speech source-based features have also been investigated [12-14]. Deep neural network-based feature extraction as well as end-to-end classification using deep neural networks has also been explored. Bottleneck features extracted from a Light Convolutional Neural Network (LCNN) have given promising results [15]. An embedding that learns within class similarities using a deep Siamese architecture was also shown to lead to improved results over many other systems [16].

Auditory based concepts such as constant-Q transform which provides time-frequency resolution similar to that of the auditory system [3] and auditory filterbanks [17] have also been used in replay attack detection systems. But, level-dependent gain adaptation of the auditory system has not been explored in this context. The current study focuses on highlighting frequency components which are low in magnitude by adaptively controlling the gain of a filterbank, thereby enhancing less noticeable channel information which could be instrumental in discriminating replayed speech samples. This technique is inspired by active mechanisms in the human cochlea.

The human ear is capable of detecting input stimuli of very low amplitude under diverse acoustic conditions. Early physiological studies on human auditory system have shown that cochlear mechanism consists of a passive as well as an active component to achieve such high sensitivity [18-20]. In particular, the selectivity and sensitivity of the cochlea actively adapts to the input stimulus level. Based on such physiological evidence, many computational and analytical models to replicate the active and passive behavior of cochlea have been developed [21-24]. Passive behavior is typically modelled as a cascaded bank of filters tuned to different frequencies [25]. Different active mechanisms have been incorporated into the passive models, making the overall filterbank behavior adaptive in nature [22].

Adaptive filterbank based feature extraction has been used in other speech processing applications. For example, it has been used in speech recognition to enhance weaker formants [24] and speech enhancement [26]. A complete model of the auditory nerve, which consists of the level-dependent adaptive mechanism, has been used in feature extraction for speaker recognition in noisy conditions [27].

http://dx.doi.org/10.21437/Interspeech.2019-1535

This study proposes the concept of level-dependent selectivity of the cochlea to improve selectivity of a second-order infinite impulse response (IIR) bandpass filterbank as a means to improve replay attack detection performance. Spatial differentiation, another concept of auditory modelling that was previously proposed by the authors [28], has also been incorporated in the current work to improve detection performance further.

2. Proposed Adaptive Filterbank Framework

The proposed adaptive filterbank framework is shown in Figure 1. As discussed before, current work aims to use adaptive filter mechanism to improve feature extraction process by making the above filterbank level-dependent. The most straightforward method to boost a low-level input stimulus is to increase the gain of a filter. Filter gain can be increased by increasing its Q factor, which in turn will improve the selectivity of the filter as well. This is further discussed in section 2.3. Each main step of the process is described in the following sections.

2.1. Bandpass Filtering

A parallel filterbank consisting of second-order IIR bandpass filters is used to decompose speech signals into subbands. Each filter in this filterbank is defined as:

$$H_{Bi}(z) = G_i \frac{1-z^{-2}}{1-2r_i\cos\theta_i z^{-1}+{r_i}^2 z^{-2}} \quad (1)$$

where i denotes the filter number and θ_i and r_i denote pole angle and pole radius of the i^{th} filter respectively. G_i is a filter specific constant given by $G_i = \frac{(1-r_i)\sqrt{1-2r_i\cos 2\theta_i+{r_i}^2}}{2\sin\theta_i}$. Time domain filtering using the above filterbank is used to first decompose a speech frame into N subbands.

2.2. Spatial Differentiation

Spatial differentiation models the mechanical coupling in the basilar membrane [29] and has the effect of increasing the order of the filters in a parallel filterbank, thereby modifying the filter characteristics. The positive effect of spatial differentiation on replay attack detection performance was discussed in our previous work [28]. The spatially differentiated output of two filters is simply the subtraction of the output of those two filters. The outputs from N bandpass filters are spatially differentiated k times and input to the adaptation system.

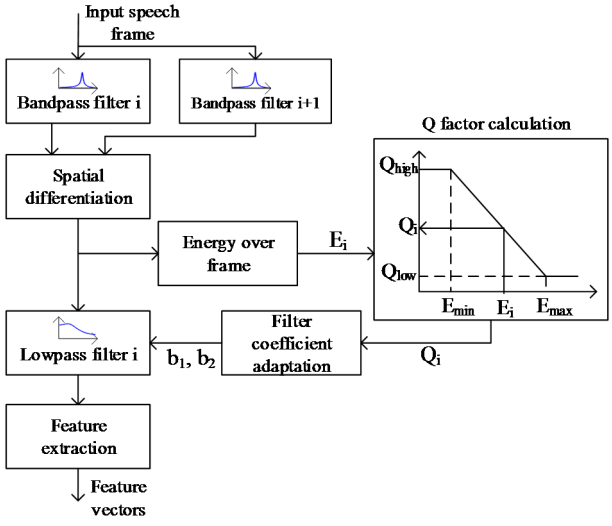


Figure 1: *Adaptive filter mechanism for the i^{th} band of the parallel filterbank.*

2.3. Adaptative-Q filtering

The spatially differentiated output of each bandpass filter is used to adjust the gain of the filterbank. Following cochlea modelling techniques, a secondary filter (tuned to the same centre frequency) with variable gain is used to make the filterbank level-dependent [22]. Filter coefficients of the secondary filter are adapted by varying the Q factor, which in turn adapts the gain of the filter. Relationship between Q factor and the gain is discussed below. The second filter used here is a second-order resonant lowpass filter with the transfer function given below:

$$H_{Li}(z) = K_i \frac{1}{1-2r_i\cos\theta_i z^{-1}+{r_i}^2 z^{-2}} \quad (2)$$

where, $K_i = 1-2r_i\cos\theta_i+{r_i}^2$ is a constant for each filter. This second-order lowpass filter shows resonant behavior at the pole frequency (resonant frequency) θ_i. Gain of the i^{th} filter at θ_i is given by:

$$|H_{Li}(\theta_i)| = \frac{1-2r_i\cos\theta_i+{r_i}^2}{(1-r_i)\sqrt{1-2r_i\cos 2\theta_i+{r_i}^2}} \quad (3)$$

Pole radius r_i at a given resonant frequency f_i $(=\theta_i\frac{f_s}{2\pi})$ of the above filter is related to the pole bandwidth BW_i by

$$r_i = 1-\pi.\frac{BW_i}{f_s} \quad (4)$$

Here, f_s is the sampling frequency. Hence, lower bandwidth gives rise to higher pole radius, which in turn increases filter gain at resonant frequency (See Equation 3). Due to the relationship $Q_i = f_i/BW_i$, where Q_i gives the Q factor of the filter,

$$r_i = 1-\pi.\frac{f_i}{Q_i f_s} \quad (5)$$

Therefore, it is seen that pole radius is proportional to the Q factor of the filter. Therefore, based on equation (3), gain of the lowpass filter can be varied by controlling its Q factor. Due to the resonant nature of the lowpass filter, a sharp gain (or selective gain) will be provided at the resonant frequency. Figure 2 shows the variation of the gain of a selected resonant lowpass filter (resonant frequency 3.3 kHz) for three Q factors.

Energy of the bandpass filtered and spatially differentiated signal is calculated over a certain time interval and used to control the Q factor of the lowpass filter (See Figure 1). The Q factor of each lowpass filter varies between two defined values. If the energy E_i is lower than a certain minimum threshold value (E_{min}), the Q factor Q_i is set to the maximum (Q_{high}). If the energy is higher than a maximum threshold value (E_{max}), the Q factor is set to the minimum value (Q_{low}). If the energy falls in between the two limits, the Q factor takes an intermediate value between the minimum and the maximum. The relationship between Q and energy is linear within this limit (See Figure 1). The new lowpass filter coefficients ($b_1 = 2r_i\cos\theta_i$ and $b_2 = {r_i}^2$ in equation (2)) are calculated based on adapted Q values and the next frame is processed using the newly adapted filters.

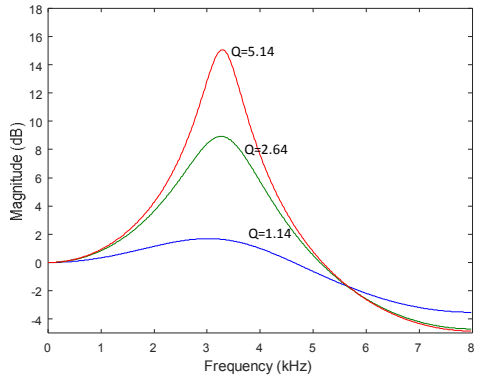
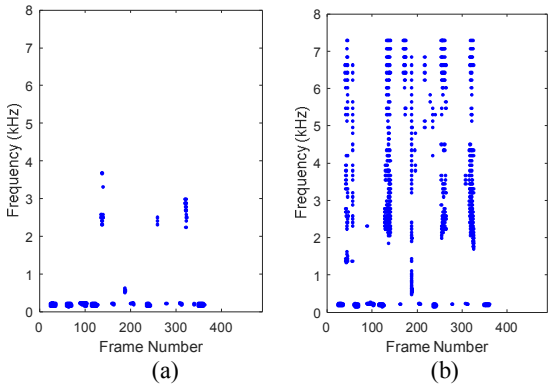


Figure 2: *Variation of gain of a resonant lowpass filter (resonant frequency 3.3 kHz) with Q factor. Increased Q factor leads to increased and more selective gain.*

Filter adaptation enhances the low magnitude components of a speech signal by increasing the Q factor. The Q_{low} limit for each lowpass filter is assigned so that the gain is maintained at 0 dB when the bandpass filter output exceeds the maximum energy threshold. Therefore, high energy components are not attenuated below the passive (no adaptation) level. The effect of adaptive filtering on a selected speech signal is shown by the pseudo-spectrograms in Figure 3. The pseudo-spectrogram plots the frequency points at which the lowpass filter output energies exceed a certain threshold. Additional peak components are seen in Figure 3(b) because the low magnitude frequency components have been boosted through adaptation. The importance of high frequency components for replay attack detection is known [10]. Therefore, enhanced high frequency components could be especially instrumental in discriminating replayed speech from genuine speech, along with other components.



(a) (b)

Figure 3: *Pseudo-spectrograms of filtered output without adaptation (3(a)) and with adaptation (3(b)).*

3. Feature Extraction

The next step is to evaluate the adaptive filterbank framework for replay spoofing attack detection. Three subband based features were extracted for the evaluation. Two of the features are the same as those used in our previous work [28], but the filtering process is adaptive instead of the conventional passive subband filtering. The speech signals are windowed into small non-overlapping frames and adaptively filtered. The resulting subband signals are full-wave rectified (taking absolute) and transformed into the frequency domain. Three features are extracted from the frequency spectrum of the full-wave rectified signals (i.e. spectral envelope of the signal). A brief description of the features used is given below.

3.1. Spectral Envelope Energy (SEE)

The first feature SEE is the average energy across each spectral envelope. SEE of k^{th} subband is calculated as given below.

$$SEE_k = \sum_{f=f_l}^{f_u} |W[f]|^2 \quad (6)$$

where f is the 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 Frequency (CF)

The CF feature is defined as the weighted average frequency of the spectral envelope of the selected frame. Here, the weights are the magnitudes of each frequency component.

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

3.3. Spectral Envelope Centroid Magnitude (CM)

The final feature CM is defined as the weighted average magnitude of the envelope under consideration. The weights are frequencies of each magnitude component.

$$CM_k = \frac{\sum_{f=f_l}^{f_u} f.|W[f]|}{\sum_{f=f_l}^{f_u} f} \quad (8)$$

4. Experimental Setup

4.1. Database

Experiments reported in this paper were conducted using the ASVspoof 2017 Version 2 database [3]. The training set of the database consists of about 1.09 hours of genuine (non-replayed) speech and about 1.03 hours of replayed speech. The evaluation set consists of about 11.10 hours of utterances in total. Therefore, the evaluation set consists of replayed utterances generated using replay conditions unseen in the training set. Speech signals are sampled at 16 kHz frequency.

4.2. Front-end Configuration

4.2.1. Filtering and Adaptation

Speech signals were windowed into non-overlapping frames of 5ms and filtered adaptively. A small frame size is required because subsequent filter coefficients of the lowpass filters are estimated based on current inputs. 80 second-order bandpass filters spaced equally in Hz frequency scale were used for subband decomposition. The Q factors of these filters increase with frequency and were kept fixed during the adaptation process. 6 spatial differentiations were applied to filtered speech signals [28]. The energy of spatially differentiated output was calculated as the sum of squares of all samples in dB scale.

All filters were first placed in non-adapted state (passive state) and subjected to adaptation which brings them to active state. The passive state Q factors (Q_{low}) of the lowpass filters linearly vary from 0.9 to 1.7. The active state maximum Q factors (Q_{high}) vary linearly from 4.9 to 5.7. It is important to

note that increasing the Q factors of filters increases the filterbank ripple. Therefore, care should be taken to ensure that the filterbank ripple is not too high as it may degrade the detection performance [30]. Minimum and maximum energy thresholds can be calculated in different ways. In this paper they were calculated by applying a sine wave to each filter with the frequency equal to the centre frequency of that filter [24]. The output average energy over a period of time was calculated, and 50 dB below the average was chosen as minimum threshold and 10 dB above average was chosen as the maximum threshold. These levels were empirically decided by applying adaptive filtering to some selected speech signals.

4.2.2. Feature Extraction

Once signals are decomposed into frequency bands, the feature extraction process is carried out. CF, CM and SEE features are extracted from subband decomposed speech frames. Since spectral envelope is of low frequency, features are obtained from frequency components up to 950 Hz. Log and DCT (discrete cosine transform) were applied to all three features to decorrelate them and first 40 coefficients were retained. Further details on feature extraction can be found in [28].

4.3. Back-end Configuration

A Gaussian Mixture Model (GMM) was used to model the data. Based on the feature dimension and the dataset size, 512 mixture component GMMs were modelled for genuine and spoofed feature distributions. The log-likelihood ratio for each test utterance was calculated during the testing phase. The MSR Toolkit [31] was used for GMM modelling.

5. Experimental Results

Equal Error Rate (EER) was used as the performance metric for the experiments. Data from training and development sets were pooled together to train the model to test on the evaluation set. Results obtained are given in Table 1. First, experiments were conducted using each feature individually. Velocity (Δ) and acceleration (ΔΔ) components of the three features were appended to include dynamic information. Next, features were fused at the feature level. Six spatial differentiations (SD) have been applied on the bandpass filter outputs in all settings unless stated otherwise.

All four selected baselines (B1, B2, B3 and B4) use GMM back-ends. Results of the first baseline (B1) is based on Version 2 of the database while results provided by B2, B3 and B4 systems are based on Version 1 of the database [33]. B1 uses Constant Q Cepstral Coefficient (CQCC) features extracted from the Constant Q Transform (CQT). This system is comparable with the proposed framework mainly because it is also auditory inspired [3]. The second baseline (B2) has utilized modulation static energy (MSE) as a front-end feature [32]. System B3 used a conventional spectral centroid magnitude (SCM) feature [9]. The B4 system is a score-level fusion of AM and FM features obtained via an auditory filterbank learnt using Convolutional Restricted Boltzmann Machine (ConvRBM) [17].

It is seen from Table 1 that the proposed CM feature-based system (S2) and SEE based system (S3) surpassed the selected single system baselines. However, S1 has shown worse performance than the baselines. The reason could be decrease in relative difference between spectral peaks in the speech signals due to adaptation, which consequently reduces the variations in CF feature. The result of feature level fusion of the single systems is shown next: S1 and S2 features together have brought down the EER possibly due to the complementary information provided by the two features. Similar behavior has been shown by S1 and S3 system features. However, concatenation of CM and SEE (S2+S3) has not shown an improvement possibly due to a correlation between the two features.

Table 1: *Comparison of experimental results (EER %) using the proposed framework with selected baselines from the ASVspoof 2017 database.*

ID	Feature	EER
Baselines		
B1	CQCC + Δ + ΔΔ [3]	12.24
B2	MSE-CC [32]	11.97
B3	SCM + Δ + ΔΔ [9]	11.49
B4	AM-ConvRBM-CC + FM-ConvRBM-CC [17]	8.89
Single systems		
S1	CF + Δ + ΔΔ (Adapted Q)	17.99
S2	CM + Δ + ΔΔ (Adapted Q)	9.42
S3	SEE + Δ + ΔΔ (Adapted Q)	10.23
Fused systems		
	S1 + S2 (Adapted Q)	**8.52**
	S1 + S3 (Adapted Q)	10.14
	S2 + S3 (Adapted Q)	9.45
	S1 + S2 (Passive, low-Q)	8.87
	S1 + S2 (Passive, high-Q)	8.93
	S1 + S2 (Adapted Q, no SD)	10.16

For comparison, systems that fused CM and CF have been tested using passive filterbanks, with all lowpass filters set to their minimum Q values and maximum Q values respectively. It is seen that the EERs are slightly higher than the adapted version, which could prove the effect of the adaptation process. However, further experiments are required to quantify the improvement. The final result given shows the positive effect of spatial differentiation on the detection results as well.

6. Conclusion

This paper proposes a novel approach inspired by cochlear modelling research for subband decomposition of speech signals to extract front-end features by adaptively adjusting gain of each filter as a function of the input signal level. This adaptive filter mechanism enhances low magnitude signal components of a speech signal actively and therefore assumed to be enhancing channel information embedded in replayed speech signals. Contrastive results provided have shown the effect of the proposed filtering technique, highlighting the potential of using classic signal processing techniques for feature extraction. The applicability of the proposed filter framework is not limited to replay attack detection. Since the main focus of the current study is to develop a better front-end framework, a typical GMM was used as the back-end. Future work on the proposed study includes incorporating deep learning frameworks to actively determine suitable Q factors for the filters based on input signal levels, and detailed study of enhanced information content provided through adaptation.

7. References

[1] 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.

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

[3] 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.

[4] 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)*, 2014, pp. 1-5: IEEE.

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

[6] 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*, 2011, vol. 4, pp. 1708-1713: IEEE.

[7] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," *Biometrics and ID Management,* pp. 274-285, 2011.

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

[9] 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.

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

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

[12] 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.

[13] 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.

[14] M. Kamble, H. Tak, and H. Patil, "Effectiveness of Speech Demodulation-Based Features for Replay Detection," *Proc. Interspeech 2018,* pp. 641-645, 2018.

[15] 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.

[16] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric," *Proc. Interspeech 2018,* pp. 671-675, 2018.

[17] 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.

[18] P. Sellick, R. Patuzzi, and B. Johnstone, "Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique," *The journal of the acoustical society of America,* vol. 72, no. 1, pp. 131-141, 1982.

[19] S. T. Neely and D. Kim, "A model for active elements in cochlear biomechanics," *The journal of the acoustical society of America,* vol. 79, no. 5, pp. 1472-1480, 1986.

[20] S. J. Elliott and C. A. Shera, "The cochlea as a smart structure," *Smart Materials and Structures,* vol. 21, no. 6, p. 064001, 2012.

[21] H. Davis, "An active process in cochlear mechanics," *Hearing research,* vol. 9, no. 1, pp. 79-90, 1983.

[22] J. M. Kates, "A time-domain digital cochlear model," *IEEE Transactions on Signal Processing,* vol. 39, no. 12, pp. 2573-2592, 1991.

[23] J. M. Kates, "Accurate tuning curves in a cochlear model," *IEEE Transactions on Speech and Audio Processing,* vol. 1, no. 4, pp. 453-462, 1993.

[24] E. Ambikairajah and L. Kilmartin, "An Adaptive Cochlear Model for Speech Recognition," in *Second European Conference on Speech Communication and Technology*, 1991.

[25] 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.

[26] D. J. Darlington and D. R. Campbell, "Sub-band adaptive filtering applied to speech enhancement," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, vol. 2, pp. 921-924: IEEE.

[27] 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.

[28] B. Wickramasinghe, E. Ambikairajah, J. Epps, V. Sethu, and H. Li, "Auditory Inspired Spatial Differentiation for Replay Spoofing Attack Detection," in *ICASSP 2019*, pp. 6011-6015, 2019.

[29] 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.

[30] J. Kua, J. Epps, E. Ambikairajah, and M. Nosratighods, "Front-end diversity in fused speaker recognition systems," *The Proceedings of APSIPA ASC,* pp. 14-17, 2010.

[31] 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, pp. 1-32, 2013.

[32] 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.

[33] T. Kinnunen *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," *submitted to Interspeech,* vol. 2017, 2017.