An adaptive transmission line cochlear model based front-end for replay attack detection

Tharshini Gunendradasan[a], Eliathamby Ambikairajah[a], Julien Epps[a], Vidhyasaharan Sethu[a], Haizhou Li[b]

[a]*University of New South Wales, Sydney, Australia*

[b]*National University of Singapore, Singapore*

Abstract

The cochlea is a remarkable spectrum analyser with desirable properties including sharp frequency tuning and level-dependent compression and the potential advantages of incorporating these characteristics in a speech processing front-end are investigated. This paper develops a framework for an active transmission line cochlear model employing adaptive notch and resonant filters. The proposed model reproduces the observed asymmetric auditory filter shape with a sharp high-frequency roll-off and level-dependent nonlinear dynamic range compression characteristics. Experimental analysis demonstrates that sharp frequency tuning and dynamic range compression of the proposed model lead to an enhanced spectral representation compared with other spectral analysis methods. The proposed model was employed in the front-end of replay spoofing attack detection systems, and experiments on the ASVspoof 2017 version 2.0 and ASVspoof 2019 databases demonstrate that the proposed model outperforms linear and nonlinear level-dependent parallel filter bank auditory models and classical spectro-temporal front-ends. The use of the proposed model leads to relative improvements of 45.6%, 51.9% and 60.8% over the baseline feature CQCCs of ASVspoof v2.0 and CQCCs and LFCCs of ASVspoof2019 evaluation datasets, respectively.

Keywords: Active cochlea model; sharp frequency tuning; dynamic range compression; level dependent nonlinearity; spoofing attack; speaker verification

1 Introduction

Auditory model front ends are integrated into a vast majority of the speech processing systems and have been shown to outperform conventional speech processing techniques (Kim et al., 1999),(Tchorz and Kollmeier, 1999). Multiple approaches to computational auditory modelling have been reported in the literature. For example, conventional auditory filters have been implemented using a set of overlapping parallel filter banks (Hohmann, 2002),(Irino and Patterson, 2006). Alternatively, transmission line auditory models (Lyon, 1997) (Kates, 1991), a cascade of digital filters that closely mimic underlying cochlea physiology have also been developed. These transmission line models reproduce auditory responses more realistically than parallel filter bank models (Lyon, 2011b),(Hemmert et al., 2004).

Sharp frequency tuning and nonlinear level dependent dynamic range compression are known to be two prominent phenomena responsible for the sensitivity and selectivity of the auditory systems over a broad intensity and frequency range (Moore, 1985),(Robles and Ruggero, 2001). Measurements of mammalian cochlea demonstrate that the cochlea has remarkable frequency selectivity with a steep high-frequency roll-off (Moore, 1978). This improved frequency selectivity in turn could lead to noise robustness (Li, 2009).

The level-dependent nonlinear dynamic range compression is achieved via an active feedback mechanism that modifies the auditory response such that low amplitude input signals are boosted. This contributes to increasing the speech intelligibility (French and Steinberg, 1947), (Villchur, 1989).Auditory models that include level-dependent nonlinearity have been shown to improve the generalisability of speech enhancement systems (Baby and Verhulst, 2018) and have been successful in analysing, classifying and recognising sounds in applications such as audio content categorisation and music recommendation (Lyon, 2011b).

A number of active auditory models that include the level dependent nonlinearities have been validated by comparing response characteristics with the available experimental measurements of the cochlea (Walters, 2011), (Kates, 1993). However, their application in different speech processing systems has not been extensively investigated thus far.

In this paper, an active cochlear model that is focused on reproducing the sharp frequency tuning and level-dependent nonlinear characteristics of the cochlea in a way that closely matches the physiological observations is developed. A front-end based on this model is then developed for replay spoofing attack detection in automatic speaker verification systems. The channel and environmental acoustic distortions are the key discriminative cues used to identify the replay attack (Wu et al., 2015), (Singh and Pati, 2019). It is anticipated that the proposed model will effectively capture these discriminative cues from regions of silence, pauses or low speech amplitude. The proposed model is an extends earlier work published by the authors (Gunendradasan et al., 2019a),(Gunendradasan et al., 2019b) to incorporate level-dependent non-linear dynamic range compression.

2 Related work

This section discusses the literature on the auditory models that incorporate sharp frequency tuning and nonlinear level-dependent cochlea characteristics as well as some background on replay spoofing attack detection.

2.1 Frequency selectivity in auditory models

Frequency selectivity of the cochlea is an essential factor for the perception of loudness, timbre, and pitch, and to understand the speech signals, particularly in a noisy environment. Experimental observations reveal that the auditory filters are asymmetric with a rounded peak and have very steep slopes with a steeper high-frequency roll-off (390dB/octave at 1kHz) compared to the low-frequency roll-off (135dB/octave at 1kHz) (Moore, 1978).

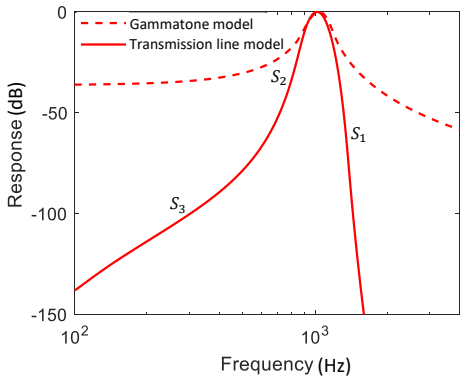

Fig. 1. Comparison of auditory filter shapes of a transmission line cochlear model (Ambikairajah et al., 1989) with the gammatone filter.

Early auditory filters used simple resonant filters to represent critical bands but suffered from poor frequency selectivity (Patterson, 1976). The findings of psychoacoustic and physiological experiments motivate the evolution of the now widely known gammatone families of filter banks (Patterson et al., 1987). The impulse response of the gammatone filters characterises the impulse response obtained at the neural nerve fibres of the cat cochlea (Johannesma, 1972). These filters have been successfully used in many speech processing applications (Qi et al., 2013),(Liu, 2018),(Yin et al., 2011). However, the magnitude response of gammatone filters does not accurately represent the shape and the sharp frequency tuning of the auditory filter response.

In most auditory models reported in the literature, the slope measurements did not receive much attention, and the selectivity of the filters are reported in terms of the quality factor (Q-factor). However, Q-factor is a more general estimate of the selectivity of the auditory filter, and it does not correctly describe the essential details of auditory tuning and the shape of the auditory filters (Allen, 2001).

The transmission line cochlear models (Ambikairajah et al., 1989) which closely model the cochlea mechanics have the inherent advantage of reproducing the observed auditory filter shapes as illustrated in Fig. 1, which is compared against gammatone auditory filter. It is clear from the figure that the transmission line cochlear model has a very sharp frequency tuning compared with a gammatone filter. It has a very steep slope S_1 on the high-frequency side and is less steep on the low-frequency side. The low side is defined by a steeper slope S_2 near the peak frequency followed by a shallower slope S_3.

2.2 Dynamic range compression in auditory models

Mammalian auditory system also have the capacity to detect and analyse a broad range of input signal intensities ranging up to 120 dB SPL. The outer hair cell (OHC) in the cochlea enable the auditory system to compress this large dynamic range in the input to a much smaller dynamic range in the neural signals by amplifying the low amplitude signals (Lyon, 1990),(Lyon and Mead, 1988). Most active cochlear models that include this level-dependent compression nonlinearity adapt the model filter parameters to achieve the desired compression. In the dynamic compressive gammachirp (dcGC) model, the filter response changes with signal level are implemented by varying the centre frequency of the asymmetric high-pass function (Irino and Patterson, 2001). Both MBPNL (Goldstein, 1990) and DRNL (Meddis et al., 2001) used similar parallel architectures and nonlinear functions to model the nonlinear psychoacoustic properties. Gammatone filters, such as all-pole gammatone filters (APGF) (Lyon, 1996), and the cascade models CAR-FAC (Lyon, 2011a) and all-pole filter cascade models (Lyon, 1997) achieved the compressive mechanism by varying the pole position. It is anticipated that the high frequency selectivity arising from the steep roll-offs of the auditory filters combined with the dynamic range compression of the auditory system benefits speech and audio perception under a range of adverse conditions. We expect these properties will also be useful in picking up channel characteristics in replay attack detection.

2.3 Replay spoofing attack

Automatic speaker verification (ASV) technology which uses voice-based authentication is now a widely adopted technology as a security measure in many applications. However, evidence shows that current ASV systems are vulnerable to malicious spoofing attacks where an unauthorised user uses an illegitimate speech sample that sounds like an authorised speaker to trick the ASV system into granting access. Thus, the prevention of malicious spoofing attacks is currently acknowledged as a priority area of investigation for the deployment of ASV systems and is an emerging field of research (Wu et al., 2015).

Spoofing attacks are categorised into four major types: replay (Gałka et al., 2015), speech synthesis (SS) (Hanilci et al., 2016), voice conversion (VC) (Kinnunen et al., 2012), and impersonation (Lau et al., 2004). A replay attack is performed by playing the pre-recorded speech of a legitimate speaker back to the ASV system. Among the four spoofing variants, replay attacks pose a significant threat to the ASV system. This is because they are the most accessible type of attack as almost everyone has access to portable even high-quality recording and playback devices such as smartphones.

Replayed speech signal can be considered as an exact copy of a genuine speech signal in terms of preserving the speech content and speaker identity, but the additional channel and environmental acoustic distortions are introduced during the recording and playback process. The presence of such distortions is, in fact, the only difference between genuine and replayed speech signals and should be explored to differentiate them.

Spectral decomposition forms the front-end of most spoofing attack detection systems. Time-frequency representation techniques such as short-time Fourier transform (STFT) (Witkowski et al., 2017), constant-Q transform (CQT) (Delgado et al., 2018) and different auditory models have been used as the front-end for this task. Auditory models employed in this context include classical parallel filterbank models that use Gabor and Butterworth filters (Kamble et al., 2018), (Kamble et al., 2020), (Kamble and Patil, 2020) and parallel filterbank models that produce sharp frequency tuning (Wickramasinghe et al., 2019b) and level-dependent compression (Wickramasinghe et al., 2019a). Experimental results reported for these models suggest that replay detection systems based on auditory models tend to be more effective than classical time-frequency representation techniques. Parallel filterbank models that incorporate high selectivity and level-dependent nonlinearities have previously been found to be useful for replay detection systems. In this paper we investigate if

the transmission line model, with greater frequency selectivity and dynamic range compression can be even more effective.

3 Proposed adaptive transmission line (ATL) cochlear model

This section presents the implementation details of the proposed active transmission line cochlear model developed from the analytical electrical representation of the cochlea. It introduces relevant background on the passive transmission line cochlear models before the proposed adaptive transmission line cochlear model is detailed.

3.1 Passive transmission line cochlear models

In the cochlea, the proposition of acoustic signals is modelled as a travelling wave that moves from the base to the apex of the basilar membrane driven by pressure variation caused by the acoustic signal in the cochlea fluid. The maximum membrane displacement occurs at the stapes (base) for high frequencies and at the far end (apex) for low frequencies. In the transmission line cochlea filter model, the basilar membrane is viewed as a cascade of small sections, with each section modelled as a set of filters (Fig. 2 represents one such "filter section"). Total number of N filter sections are then cascaded to represent the whole cochlea. The passive transmission line cochlear model proposed in (Ambikairajah et al., 1989) determine the filters that form the filter section based on the transfer functions derived from the transmission line electrical representation of the passive cochlea. This model was developed to produce linear cochlea characteristics focusing on reproducing measured auditory filter shapes and frequency tuning. The isolated electrical section representing a small part of the basilar membrane was used to derive the transfer functions. The section is separated from its neighbours by loading it with the input impedance from the remaining sections.

The isolated electrical equivalent circuit of the passive cochlea will look similar to Fig. 2 (a) if the time-varying voltage source $V_{OHC}(s)$ is suppressed (Ambikairajah et al., 1989). The impedances R, L and C represent the electrical impedances of the basilar membrane. The input voltage $V_i(s)$ and the output voltage $V_o(s)$ represent the input and output pressures in the particular cochlear filter section and the voltage across the capacitor $V_m(s)$ represents the displacement of the basilar membrane. $V_{Th}(s)$ and Z_{Th} represent the Thévenin voltage and impedance obtained during the isolation process (Ambikairajah et al., 1989). The electrical circuit model of a filter section, Fig. 2 (a), can equivalently be modelled as a cascade of lowpass, resonant and notch filters as illustrated in Fig. 2 (b) when feedback via outer hair cells are not considered.

3.2 Proposed ATL cochlear model

In the well-known electrical transmission line model, the effect of the outer hair cells are introduced as a voltage sources (Giguere and Woodland, 1994). Here we propose that this voltage source be modelled as a dependent voltage to introduce the desired feedback which in turn will provide the dynamic range compression. Specifically, as shown in Fig. 2 (a), we assumed that $V_{OHC}(s)$ is proportional to $V_{Th}(s)$,

$$V_{OHC}(s) = \tilde{g} V_{Th}(s) \tag{1}$$

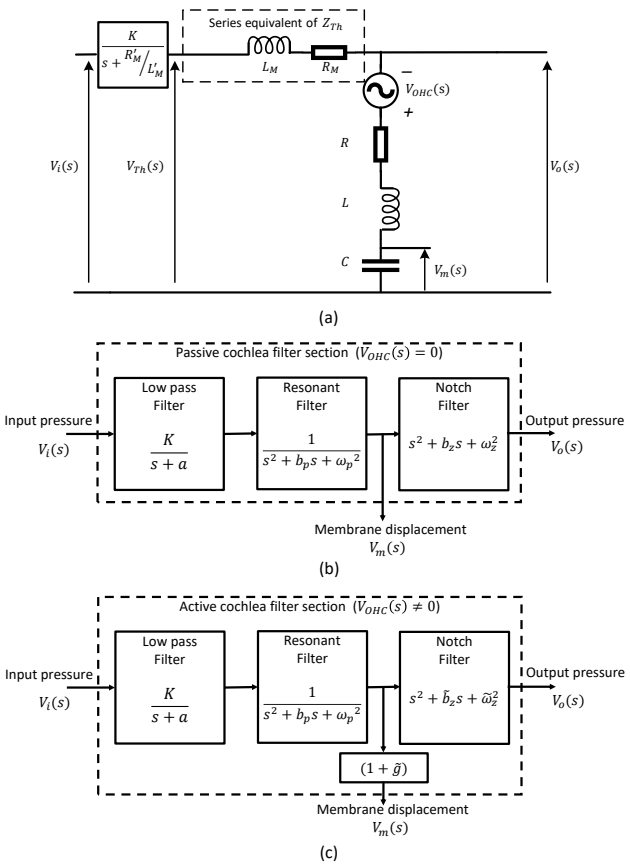

Fig. 2. (a) Proposed simplified electrical equivalent circuit of an isolated single section of the active cochlea. If the voltage source $V_{OHC}(s)$ is eliminated it will represent the passive cochlea (Ambikairajah et al., 1989) (b) The equivalent filter representation of passive cochlea when $V_{OHC}(s) = 0$ in (a). (c) The equivalent filter representation of active cochlea depicted in (a). Here $\tilde{g}$, $\tilde{b}_z$ and $\tilde{\omega}_z$ represent adaptive parameters.

where $\tilde{g}$ is an adaptive gain that varies with time (note that in this paper we use $\tilde{}$ to denote adaptive parameters). We also further simplify the circuit by replacing the Thévenin impedance Z_{Th}, previously represented as a parallel connection of resistor R'_M and inductor L'_M (Ambikairajah et al., 1989), with its series equivalent resistance R_M and inductance L_M as illustrated in Fig. 2 (a).

From the proposed isolated circuit, the pressure and displacement transfer functions which are both essential to describe the wave propagation in the cochlea are derived. The pressure transfer function, $P(s) = V_o(s)/V_i(s)$, models how the pressure wave travels along the basilar membrane,

$$P(s) = \frac{K}{s+a} . \frac{1}{s^2 + b_p s + {\omega_p}^2} . (s^2 + \tilde{b}_z s + \tilde{\omega}_z^2), \tag{2}$$

where $\tilde{\omega}_z$ is the adaptive notch frequency $\tilde{\omega}_z = \sqrt{1/(C(L - \tilde{g} L_M))}$, $\tilde{b}_z = \tilde{\omega}_z/\tilde{Q}_z = (R - \tilde{g} R_M)/(L - \tilde{g}\, L_M)$ where $\tilde{Q}_z$ is the adaptive Q-factor of the notch filter, ω_p is the resonant frequency $\omega_p = \sqrt{1/(C(L + L_M))}$, $b_p = \omega_p/Q_p = (R + R_M)/(L + L_M)$ where Q_p is the Q-factor of resonant filter, $a = R'_M/L'_M$ is the lowpass filter coefficient, and K is a constant. As per Eq. (2), pressure transformation can be modelled as a filter section consists of a cascade of a first order lowpass filter, a second

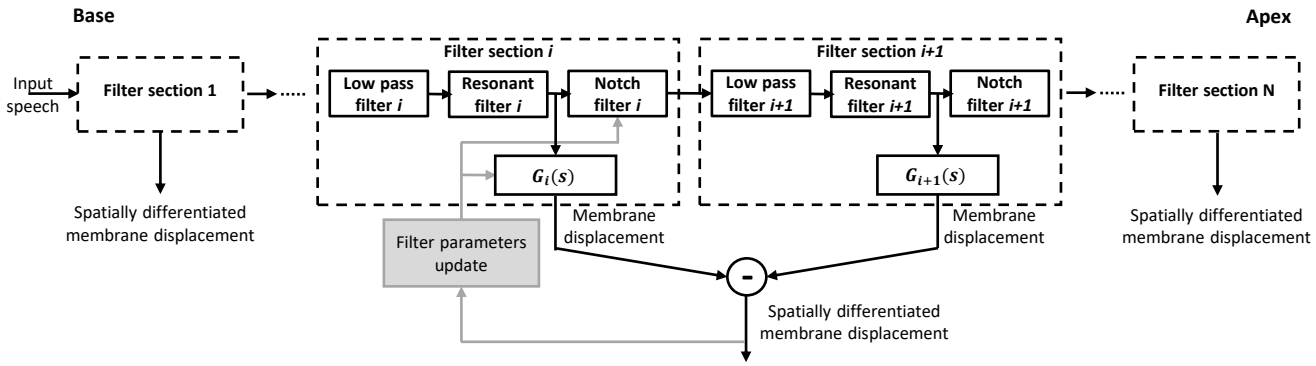

Fig. 3. Proposed ATL cochlear model. The cascaded filter sections consist of low pass, resonant and notch filters, which model the pressure along the membrane, and the output taped after the resonant filters is fed into a resonant filter $G(s)$ for membrane displacement. Membrane displacement is then spatially differentiated to increase the frequency selectivity. The active feedback that adapts the filter parameters is illustrated in gray color.

order resonant filter and an adaptive second order notch filter, as illustrated in Fig. 2 (c). The displacement transfer function, $D(s) = V_m(s)/V_i(s)$, describes the displacement of the basilar membrane at the position corresponding to that section,

$$D(s) = (1 + \tilde{g}).\frac{K}{s+a}.\frac{1}{s^2+b_p s+{\omega_p}^2}. \quad (3)$$

Since the lowpass and resonant filters are common for both pressure and displacement transfer functions according to Eq. (2) and Eq. (3), a simple combined model, as shown in Fig. 2 (c), can be implemented.

The model of the filter section in Fig. 2 (c) that represents a small part of the cochlea is then cascaded to represent the whole cochlea as illustrated in Fig. 3. The input to each filter section is the pressure and then the output pressure from that section is passed on to the following filter section (and so on). This way the pressure wave travels along the basilar membrane. The membrane displacement is tapped at the intermediate point of each filter section. In both transfer functions, given by Eq. (2) and (3), some of the filter parameters are adaptive, which in turn allows for the filter gain and the selectivity of the filter to be varied in response to the input signal characteristics. In the pressure transfer function, the notch filter parameters are adaptive. An additional adaptive gain term $(1 + \tilde{g})$ is present in the displacement transfer function to control the movement of the membrane. Here, we propose to replace, $(1 + \tilde{g})$, in the displacement transfer function with the controllable gain resonant filter $G(s)$ that controls both the gain and selectivity,

$$G(s) = \frac{{\omega_r}^2}{s^2+\tilde{b}_r s+{\omega_r}^2}, \quad (4)$$

where $\tilde{b}_r = \omega_r/\tilde{Q}_r$, ω_r and $\tilde{Q}_r$ are the resonant frequency and the adaptive Q-factor of the filter, respectively and ω_r is chosen to be equal to ω_p. It can also easily be shown that the gain of this filter at its resonant frequency ω_r is equal to its Q-factor and the gain and selectivity can be controlled by changing its Q-factor.

There are additional auditory mechanisms that are not included in the electrical equivalent circuit (Fig. 2 (a)), which further generate sharp frequency tuning of the membrane, e.g. the longitudinal shear force along the membrane (Hall, 1977). Spatial differentiation (Hall, 1977) is introduced as an additional sharpening mechanism to generate the required frequency tuning. The adjacent resonant filter $G(s)$outputs are subtracted along the length of the basilar membrane, as illustrated in Fig. 3, to get the final differentiated displacement. For convenience, only a first-order spatial differentiation is shown in Fig. 3, but two orders of spatial differentiation are applied by repeating the same process on the first order spatially differentiated output.

The level-dependent characteristics can be introduced into the transmission line cochlear models in many ways including the use of controllable Q-factor resonant filters as additional filters connected in parallel to the cascaded filters (Hirahara and Komakine, 1989). Further by changing the Q-factors of the resonant filters in the cascaded filters itself (Lyon, 2011a). In the proposed cochlear model, both the additional resonant filters and cascaded notch filter parameters were adapted to bring the nonlinearity. In particular, the Q-factor and the resonant frequency of the cascaded notch filters were changed (Eq. (2)) in contrast to many other active transmission line cochlear models that vary the Q-factor of the resonant filters alone. We refer to our proposed cochlear model as an adaptive transmission line (ATL) cochlear model.

3.3 Updating filter parameters in the proposed ATL cochlear model

In the proposed ATL model, as illustrated in Fig. 3, the filter parameters are adaptively updated based on energy of the membrane displacement outputs. Specifically, once every frame (e.g. 1ms frames), the average energy of the displacement output of each filter section is computed and the peak average energy value, $E(n)$, across k filter sections on either side of each filter section is chosen to determine the updated Q-factor, $Q_r(n)$, for the adaptive resonant filter, $G(s)$, in that section. Similarly, the Q-factor, Q_z, and notch frequency, ω_z, of the adaptive notch filter is also updated based on $E(n)$.

These filter parameters are tuned in a way that complies with the cochlea's experimental observations, where for low energy signal, the gain and selectivity of the auditory response increases whereas for high energy signal both reduce (Johnstone et al., 1986). To achieve this, Q_r and ω_z should decrease with increasing $E(n)$, while Q_z should increase with increasing $E(n)$. We proposed that $Q_r(n)$, the Q-factor of $G(s)$, varies between its maximum ($Q_{r,max}$) and minimum ($Q_{r,min}$) values linearly based on the signal energy $E(n)$, as illustrated in Fig. 4. Here $E_{Th,min}$ and $E_{Th,max}$ are the minimum and maximum energy thresholds (in dB

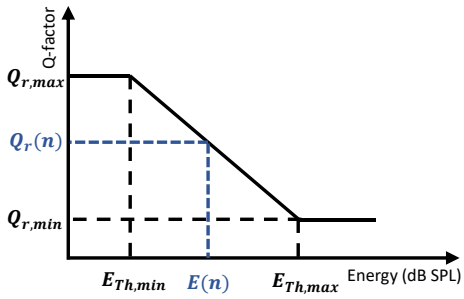

Fig. 5. Proposed adaptive Q-factor, $Q_r(n)$, of resonant filter $G(s)$ as a function of displacement output energy $E(n)$.

SPL) within which Q_r varies (refer Fig. 4) inversely proportional to $E(n)$. If the output signal energy is below $E_{Th,min}$, the Q-factor is set at its maximum value while if the energy is above $E_{Th,max}$ it is set at the minimum Q-factor. Similarly, the notch frequency $\omega_z(n)$ also varies linearly between the minimum $\omega_{z,min}$ and the maximum $\omega_{z,max}$ notch frequencies, inversely proportional to $E(n)$. The notch filter Q-factor, $Q_z(n)$, is varies between maximum $Q_{z,max}$ and minimum $Q_{z,min}$, directly proportional to $E(n)$.

To ascertain $E_{Th,min}$ for each filter section, we provide a sinusoid with a frequency matching the resonant frequency of the corresponding $G(s)$ and amplitude corresponding to 0 dB SPL at the input of the model. Then setting $E_{Th,min}$ and $E_{Th,max}$ as 0 dB SPL and 100 dB SPL respectively, we estimate the energy of the membrane displacement output of that filter section which is then taken to be the actual $E_{Th,min}$. Similarly, to ascertain $E_{Th,max}$ we set the input amplitude to correspond to 100 dB SPL and carry out the same process. Note that when the proposed model is used in a speech processing system, we estimate these threshold energies based on the minimum and maximum signal energies in the database instead of 0 dB SPL and 100 dB SPL. The values of $Q_{r,min}$ and $Q_{r,max}$ are obtained as explained in the following section.

3.4 Selection of Filter Parameters

The proposed ATL model was designed for a sampling rate of 16 kHz, and a total of 128 filter sections was used to model the cochlea. Since the proposed ATL model's intended application is replay spoofing attack detection, a linear frequency scale, which is more effective for replay detection (Font et al., 2017) , was selected to model the cochlea. i.e., the resonant frequencies (ω_p and ω_r) in each section are equally spaced between 50Hz and 7900 Hz for consecutive filter sections.

We estimate the Q-factors of the notch and resonance filters in the pressure transfer function by assuming that the Q-factor is inversely proportional to the width of the basilar membrane. The width, W, of the membrane along its length, $0 \leq x \leq 3.5cm$, is given as below:

$$W(x) = 0.019 + 0.0093x \tag{5}$$

Thus, the Q-factor of the resonant filters, Q_p, is,

$$Q_p = \frac{k_p}{W(x)}. \tag{6}$$

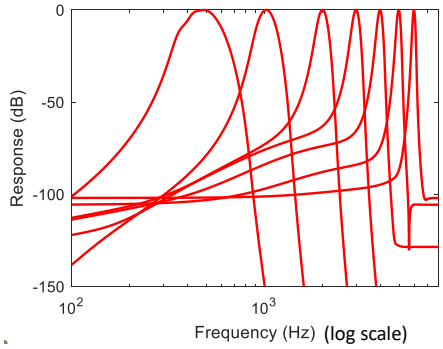

Fig. 4. Comparison of a set of frequency response curves of the proposed ATL model at around 0.5, 1, 2, 3, 4, 5 and 6 kHz. It produces the general trend of border curves at low frequencies and narrow curves at high frequencies.

where, k_p is a constant. Similarly, the Q-factor of each notch filter is given as $Q_z = k_z/W(x)$, where k_z is another constant. Based on experimental measurements of the membrane selectivity, Q_p has been measured as 8.99 and 3.31 at the basal and apical end of the basilar membrane respectively, and Q_z was chosen to be 19.73 and 7.25 at these two ends (Ambikairajah et al., 1989). These values then allow us to calculate k_p and k_z as 0.17099 and 0.37485, respectively. Finally, the filter sections are mapped to the Q-factors via x based on the Bekesy frequency scale (Olson et al., 2012). All other filter parameters are determined as per (Ambikairajah et al., 1989).

4 Proposed ATL cochlear model characteristics

The proposed ATL model produces an auditory filter shape similar to the one shown in Fig. 1 in close agreement with the mammalian cochlea's physiological tuning curves. The auditory response of the proposed model at different frequency positions are illustrated in Fig. 5. The model exhibits the desired characteristics of having broader tuning curves in the low-frequency side, whereas narrow tuning in the high-frequency side (Robles and Ruggero, 2001). A comparison of the high-frequency side slope S_1 and low frequency side slopes S_2 and S_3 (shown in Fig. 1) of the curves given in Fig. 5 with the available experimental measurements of the human cochlea are reported in Table 1. As observed in human cochleae, the proposed model exhibits much sharper frequency tuning in the high-frequency regions when compared to the low-frequency regions, which is in line with the experimental measurements reported on the human cochlea (Moore, 1978). Further, the slope measurements of the model comply well with the experimental measurements up to 4 kHz.

To analyse the selectivity of the proposed model, the correlation coefficients between the short-term energies of the basilar membrane displacement outputs of each pair of filter sections was estimated. Similarly, the correlation coefficients were calculated for the gammatone filters as well for comparison. Fig. 6 illustrates the heatmap corresponding to the average correlation matrix obtained using a large set of speech signals (the entire training partition of the ASVspoof 2017 2.0 dataset) for both the ATL and the gammatone models. The gammatone model has a relatively high correlation between the adjacent filters. By contrast, the proposed ATL model has a lower correlation reflected in the small off-diagonal elements, suggesting that the proposed ATL model leads to less overlap between frequency bands.

Table 1. Comparing the high-frequency roll-off (S_1) and low-frequency roll-offs (S_2 and S_3) of the frequency response curves of the proposed ATL model illustrated in Fig. 5 with the experimental measurements reported a forward masking experiment (Moore, 1978), (Nelson, 1991). It should be noted that measures in the forward masking experiment were carried out only on two or three humans and the measurement accuracy is around 12% to 15%.

Frequency (kHz)	High-frequency roll-off (S1) (dB/oct)		Low-frequency roll-off (S2) (dB/oct)		Low-frequency roll-off (S3) (dB/oct)	
	ATL model	human	ATL model	human	ATL model	Human
0.5	232	97-190	81	50-55	31	-
1	325	310-650	120	90-180	25	20-30
2	449	330-1820	142	84-160	19	-
3	529	-	169	-	12	-
4	622	640-2800	207	83-230	9	-
5	744	-	283	-	8	-
6	825	420-590	425	31-150	9	-

The active feedback mechanism in the proposed ATL model tunes the frequency response depending on the input signal amplitude. Fig. 7 illustrates the variations in the frequency tuning of the 1 kHz filter section when a 1 kHz sinusoidal signal of different energy levels between 0 to 100 dB SPL is passed into the model. At the lowest energy level of 0 dB SPL, the response is tuned to provide the highest gain of 33 dB to boost the energy of the signal, and it also becomes more selective. As the signal level increases, the filter gain starts to reduce, dropping to 0 dB at 100 dB SPL. Along with this frequency selectivity also becomes broader.

The input-output relationship with the varying input signal level of the proposed ATL model that is designed to apply compression between 0 to 100 dB SPL is illustrated in Fig. 8. For input signal energies ranging from 0 to 130 dB SPL, the model performs compression up to around input level of 100 dB SPL, after which it acts linearly. The amount of compression it applies varies based on the frequency. At a low frequency of 0.5kHz, it offers compression of around 36dB up to input level of 100 dB SPL. Meanwhile, at high frequency 6 kHz, the compression increases to around 52 dB, which accounts for a large compression amount. The differences in the amount of compression across low and high frequencies are in good qualitative agreement with those measured experimentally (Robles and Ruggero, 2001).

To visualise the advantage of using the proposed active cochlear model over the passive auditory model, time-frequency representations of both genuine and its corresponding replayed speech signals are illustrated in Fig. 9. Here a passive transmission line cochlear model proposed in (Ambikairajah et al., 1989) that does not include any level-dependent nonlinear compression is considered. It can be observed that in the ATL model output, low energy spectral information is emphasised to a greater degree when compared to the passive model. Therefore, ATL model highlights the differences between genuine and replayed speech signals more than the passive model.

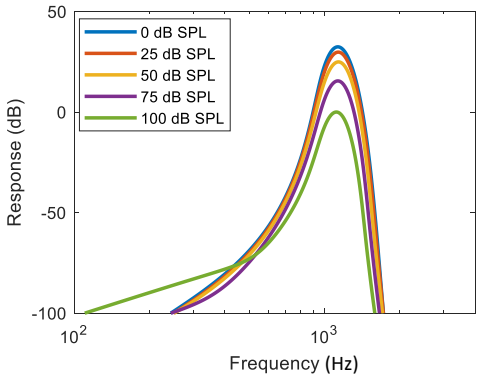

Fig. 7. The changes in frequency response of membrane displacement of the proposed ATL model at the filter section corresponding to 1 kHz for the sinusoidal signals of the same frequency, with input signal amplitude varying from 0 to 100 dB SPL in four steps 0, 30, 60, 100 dB SPL. When the input signal amplitude drops from higher to lower value, the membrane response becomes sharply tuned, and the gain increases.

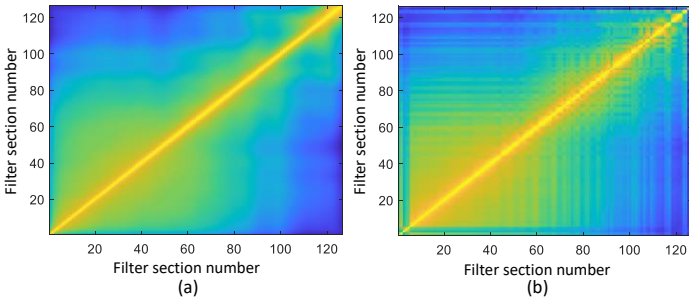

Fig. 6. Heatmaps corresponding to the average pairwise correlation coefficients between filter sections for: (a) the proposed ATL model and (b) the gammatone filter bank model. The ATL model has a more prominent diagonal with a lower correlation between the adjacent filter sections in comparison with the gammatone filter banks.

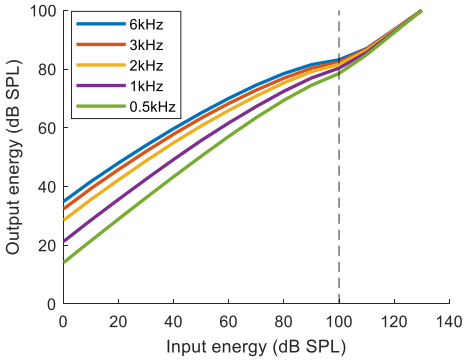

Fig. 8. Gain of the ATL model at different frequencies, with input levels ranging from 0 dB SPL to 130 dB SPL in steps of 10dB. Beyond 100 dB SPL input level, the gain is relatively constant.

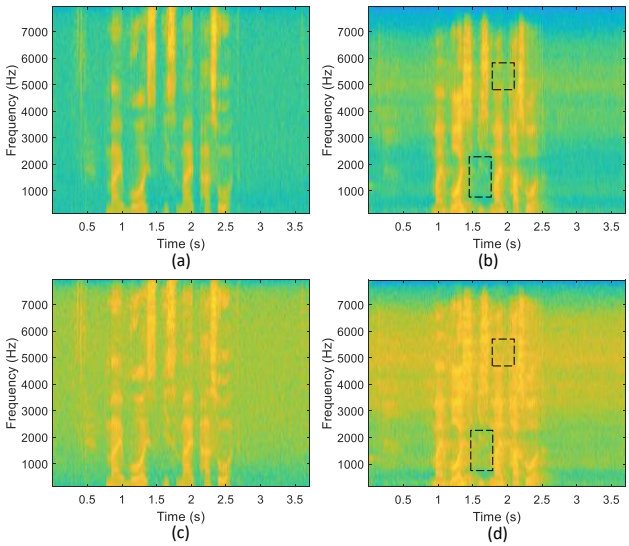

Fig. 9. Visualization of time-frequency representations of proposed ATL model and passive transmission line cochlear model (Ambikairajah et al., 1989) for genuine and replayed speech signals. (a) Passive model-genuine speech, (b) passive model- replayed speech, (c) ATL model-genuine speech, (d) ATL model-replayed speech. Low energy regions become more visible in the ATL model than the passive model due to dynamic range compression. The black rectangular boxes denote some of the low energy regions of replayed speech signals. Here the difference between genuine and replayed speech signals are more emphasized in ATL model than passive model.

5 Experimental setup

Experiments were conducted to investigate the potential benefits of the proposed ATL cochlear model as a front-end for replay spoofing attack detection. This section details the feature extraction process from the ATL model for replay attack detection. Further, the database used for the experiments, the experimental settings and the baseline model used for the comparison are discussed.

The amplitude modulation (AM) feature that tracks the amplitude envelope of the speech signal was investigated for replay detection. During replay attacks, the amplitude envelope may be distorted by the channel effects of audio recording and playback devices, noise and reverberation due to the environmental acoustics. Therefore, the amplitude modulation of the speech signal, which is considered to be important information for speech perception (Mitra et al., 2012), is explored for replay detection. For feature extraction, average output magnitude estimated within small overlapping windows was taken as the AM feature. Log compression and the DCT were then applied to the extracted AM components, to compress and decorrelate the extracted features. Along with these DCT static coefficients, the delta and delta-delta coefficients were also appended to the feature representation.

All experiments were conducted on the ASVspoof 2017 version 2.0 (Delgado et al., 2018) and ASVspoof 2019 replay databases (Todisco et al., 2019). These databases are a collection of genuine and replayed speech samples constructed for the purpose of developing replay attack countermeasures to protect automatic speaker verification systems. Replayed speech signals in ASVspoof 2017 database were generated by replaying the original utterances through various reply configurations consisting of different recording devices, loudspeakers and diverse acoustic environments. In comparison, replayed speech signals were simulated in the ASVspoof 2019 database. The detailed description of these databases are summarized in Table 2 and Table 3. The speech signals in both the databases were sampled at 16 kHz.

In the proposed ATL model, the adaptive filter parameters were update using 1 ms frames (once per frame). Eight adjacent filter sections on either side of the filter section were considered when choosing the peak average energy.

According to studies on replay detection, the high-frequency regions contain more discriminative information than low and mid frequencies (Gunendradasan et al., 2018; Witkowski et al., 2017) and consequently a 1st order FIR filter ($1 - 0.97z^{-1}$) was used for pre-emphasis. The AM features were extracted every 2ms. Cepstral mean and variance normalization was carried out on the extracted AM features.

The Gaussian mixture model (GMM) is a well-known and commonly applied classifier for spoofing attack detection. Thus, it is used as the back-end classifier to the AM features. Two 512 mixture GMM models for genuine and spoofed classes were trained using the training dataset. The log-likelihood ratio of the genuine and spoofed models was used to calculate the classification scores for testing. All the features presented in this paper for comparison purposes uses GMM as the back end classifier.

The active feedback in the ATL model introduces dynamic range compression and the proposed model was tested both with and without the active feedback to investigate the advantages of the model's nonlinear active compression property for spoofing detection. Two states of the ATL model without the active feedback was considered, namely, a "low Q" state and a "high Q" state. The "low Q" state represents the model parameter setting when the minimum Q-factor is assigned to the resonant filter $G(s)$ ($Q_r = Q_{r,min}$). Similarly, the "high Q" state is when $Q_r = Q_{r,max}$. When the adaptation is turned on with the active feedback, the filter parameters changed between these two extreme values.

The gammatone filter model, which is the popular choice to represent the impulse response of the cochlea, was used as the baseline (passive) model to compare and quantify the advantage of the high selectivity and dynamic range compression properties

Table 2. ASVspoof 2017 version 2.0 database training and test partitions.

Subset	# Replay configurations	# Utterances	
		genuine	spoof
Training	3	1507	1507
Development	10	760	950
Evaluation	57	1298	12008

Table 3. ASVspoof 2019 replay database training and test partitions.

Subset	# Utterances	
	genuine	spoof
Training	5400	48600
Development	5400	24300
Evaluation	18090	116640

Table 4. Comparison of replay detection performance based on ASVspoof 2017 version 2.0 database.

	Features	% EER	
		Development	Evaluation
Without dynamic range compression	**ATL (without feedback) – "Low Q" state**	**7.28**	**9.56**
	ATL (without feedback) – "High Q" state	**7.01**	**9.21**
	Gammatone Model	8.78	14.63
	CM (Wickramasinghe et al., 2019b)	-	10.93
	ESA-IACC (Kamble and Patil, 2020)	7.99	13.45
	CQCC (19E-SDA) (Delgado et al., 2018)	9.06	13.74
	LFCC	10.31	15.73
	MFCC	24.19	26.90
	SCMC	11.01	15.67
With dynamic range compression	**ATL (with feedback)**	**5.69**	**7.47**
	CM -Adaptive Q (Wickramasinghe et al., 2019a)	-	9.42
	SEE -Adaptive Q (Wickramasinghe et al., 2019a)	-	10.23

achieved by the proposed ATL model. The gammatone filterbank was implemented using the same frequency scale, the number of filters, and bandwidths as the ATL model, in order to aid comparison. The bandwidth of the proposed ATL model was measured and then used to estimate the Q-factors of the gammatone filters, in order to match the bandwidths of the two models.

6 Results and discussion on replay spoofing attack detection

In this section, comparisons of the proposed ATL model with other auditory models and spectral feature extraction techniques are presented, based on the AS spoof 2017 version 2.0 and ASVspoof 2019 databases. AM and short-term spectral energy based features are among the most widely used features for distinguishing genuine speech from replayed speech. The ASVspoof 2017 challenge baseline feature constant-Q cepstral coefficients (CQCC) uses CQT transform for spectral decomposition. There are other short term energy features used for replay detection that are extracted from the magnitude spectrum obtained using STFT, such as Mel frequency cepstral coefficients (MFCC), linear frequency cepstral coefficient (LFCC) and spectral centroid magnitude coefficients (SCMC).

Table 5. Replay detection performance of the proposed ATL model for different environments, playback and recording devices in terms of % EER, on the ASV spoof 2017 version 2.0 evaluation dataset. The performance obtained using baseline CQCC features are reported within parentheses for comparison.

Replay configurations	Conditions		
	Low	Medium	High
Environment	**3.36** (16.68)	**6.46** (18.73)	**8.01** (21.86)
Playback	**6.06** (16.64)	**5.92** (16.44)	**8.27** (18.37)
Recording	**6.30** (10.80)	**6.56** (15.99)	**8.02** (17.77)

In addition to the CQT and STFT time-frequency representation techniques, AM features extracted from the auditory models that use time-domain parallel filter banks have been proposed for spoofing detection, which are used to compare the performance of the proposed ATL model. The ESA-IACC feature used Gabor filters for spectral decomposition (Kamble and Patil, 2020). The CM feature proposed in (Wickramasinghe et al., 2019b) uses second-order infinite impulse response (IIR) bandpass filters and focuses on increasing the frequency selectivity of auditory filters using spatial differentiation. Extended versions of CM, CM-Adaptive-Q and SEE-Adaptive-Q (Wickramasinghe et al., 2019a), incorporated dynamic range compression into a parallel filter bank model.

Table 4 summarises the results on ASVspoof 2017 version 2.0 database for the AM features extracted from the proposed ATL model and other parallel filter bank auditory models. As explained in section 5, the proposed model was evaluated with and without active feedback to analyse the significance of nonlinear level-dependent compression for replay detection. It can be noticed that even without dynamic range compression, the ATL model outperforms all other methods suggesting that high frequency-selectivity benefits replay detection. The performances were similar for the ATL model at both the "low Q" and "high Q" states.

The proposed ATL with the dynamic range compression markedly improves performance over the configuration without its active feedback. This suggest that low energy regions contain discriminative channel and acoustic information, and boosting them can contribute toward improving replay detection performance. Moreover, the feature extracted from the proposed ATL model perform better than the CM-Adaptive-Q and SEE-Adaptive-Q features extracted from the parallel filter bank auditory model, which also incorporate the dynamic range compression, highlighting the benefit from large amount of compression and greater frequency selectivity of the proposed model.

Table 5 reports a breakdown of replay detection performance in terms of low, medium, and high threat conditions defined in (Delgado et al., 2018) for the ASVspoof 2017 version 2.0 database.

Table 6. Comparison of replay detection performance based on ASVspoof 2019 version 2.0 database.

Feature	% EER	
	Development	Evaluation
CQCC (Todisco et al., 2019)	9.87	11.04
LFCC (Todisco et al., 2019)	11.96	13.54
Gammatone Model	10.54	12.12
ATL (with feedback)	4.08	5.31

The proposed ATL model outperforms the baseline CQCC features under all settings. It can be noted that replay attacks with the noise-free acoustic environments, which are considered high threat and challenging to detect due to lack of artifacts from playback and recording devices (Delgado et al., 2018), (Kamble and Patil, 2020) are also identified with less difficulty when using the proposed ATL model.

Table 6 reports the replay detection performance as evaluated on the development and evaluations sets of the ASVspoof 2019 replay database. The ASVspoof 2019 challenge included two features, CQCCs and LFCCs, with a GMM classifier based backend as the baseline systems. Along with these two features, AM features extracted from Gammatone filter outputs are compared with the proposed ATL model. The ATL model outperforms all three features on both the development and evaluation sets. These results reinforce the superiority of the proposed ATL model over traditional front-ends in the context of replay detection systems.

7 Conclusion

This paper presents an adaptive transmission line (ATL) cochlear model that includes novel adaptive notch and resonant filters to mimic the feedback provided by outer hair cells in the cochlea. This in turn leads to a cochlear model with auditory filter shapes, frequency selectivity, and nonlinear level dependent dynamic range compression characteristics in close agreement with experimental measurements of the human cochlea. Our results show that the high selectivity achieved by the proposed ATL model contributes to improving the replay detection performance compared to the parallel filter bank auditory models such as gammatone filters. As a result of the ATL model's nonlinear dynamic range compression, low energy regions that may not be captured by most front-ends are emphasized by the proposed model, leading to a better representation of low amplitude channel characteristics and consequently aiding replay detection. Adaptive, biologically inspired auditory front-ends may also be advantageous in other areas of speech processing.

Acknowledgement

This work was funded by ARC Discovery Grant DP190102479. The authors would also like to thank the reviewers for the invaluable feedback which helped improve this paper.

References

Allen, J., 2001. Nonlinear cochlear signal processing, Physiology of the Ear, Second Edition. Singular Thompson, pp. 393-442.

Ambikairajah, E., Black, N.D., Linggard, R., 1989. Digital filter simulation of the basilar membrane. Computer Speech and Language 3, 105-118.

Baby, D., Verhulst, S., 2018. Biophysically-inspired features improve the generalizability of neural network-based speech enhancement systems, 19th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH). ISCA, pp. 3264-3268.

Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., Yamagishi, J., 2018. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements, Proc. Odyssey The Speaker and Language Recognition Workshop, pp. 296-303.

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

French, N.R., Steinberg, J.C., 1947. Factors governing the intelligibility of speech sounds. The journal of the Acoustical society of America 19, 90-119.

Gałka, J., Grzywacz, M., Samborski, R., 2015. Playback attack detection for text-dependent speaker verification over telephone channels. Speech Communication 67, 143-153.

Giguere, C., Woodland, P.C., 1994. A computational model of the auditory periphery for speech and hearing research. I. Ascending path. The Journal of the Acoustical Society of America 95, 331-342.

Goldstein, J.L., 1990. Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering. Hearing research 49, 39-60.

Gunendradasan, T., Ambikairajah, E., Epps, J., Li, H., 2019a. An Adaptive-Q Cochlear Model for Replay Spoofing Detection. Proc. Interspeech, 2918-2922.

Gunendradasan, T., Irtza, S., Ambikairajah, E., Epps, J., 2019b. Transmission Line Cochlear Model Based AM-FM Features for Replay Attack Detection, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6136-6140.

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

Hall, J., 1977. Spatial differentiation as an auditory''second filter'': Assessment on a nonlinear model of the basilar membrane. The Journal of the Acoustical Society of America 61, 520-524.

Hanilci, C., Kinnunen, T., Sahidullah, M., Sizov, A., 2016. Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. Speech Communication 85, 83-97.

Hemmert, W., Holmberg, M., Gelbart, D., 2004. Auditory-based automatic speech recognition, ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing.

Hirahara, T., Komakine, T., 1989. A computational cochlear nonlinear preprocessing model with adaptive Q circuits, International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 496-499.

Hohmann, V., 2002. Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica 88, 433-442.

Irino, T., Patterson, R.D., 2001. A compressive gammachirp auditory filter for both physiological and psychophysical data. The Journal of the Acoustical Society of America 109, 2008-2022.

Irino, T., Patterson, R.D., 2006. A dynamic compressive gammachirp auditory filterbank. IEEE transactions on audio, speech, and language processing 14, 2222-2232.

Johannesma, P., 1972. The pre-response stimulus ensemble of neurons in the cochlear nucleus, Symposium on Hearing Theory. IPO.

Johnstone, B., Patuzzi, R., Yates, G., 1986. Basilar membrane measurements and the travelling wave. Hearing research 22, 147-153.

Kamble, M., Tak, H., Patil, H., 2018. Effectiveness of Speech Demodulation-Based Features for Replay Detection. Proc. Interspeech, 641-645.

Kamble, M.R., Patil, H.A., 2020. Combination of Amplitude and Frequency Modulation Features for Presentation Attack Detection. Journal of Signal Processing Systems, 1-15.

Kamble, M.R., Tak, H., Patil, H.A., 2020. Amplitude and Frequency Modulation-based features for detection of replay Spoof Speech. Speech Communication 125, 114-127.

Kates, J.M., 1991. A time-domain digital cochlear model. IEEE Transactions on Signal Processing 39, 2573-2592.

Kates, J.M., 1993. Accurate tuning curves in a cochlear model. IEEE Transactions on Speech and Audio Processing 1, 453-462.

Kim, D.-S., Lee, S.-Y., Kil, R.M., 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on speech and audio processing 7, 55-69.

Kinnunen, T., Wu, Z.-Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4401-4404.

Lau, Y.W., Wagner, M., Tran, D., 2004. Vulnerability of speaker verification to voice mimicking, Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing. IEEE, pp. 145-148.
Li, Q., 2009. An auditory-based transfrom for audio signal processing, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, pp. 181-184.
Liu, G.K., 2018. Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech. arXiv preprint arXiv:1806.09010.
Lyon, R.F., 1990. Automatic gain control in cochlear mechanics, The mechanics and biophysics of hearing. Springer, pp. 395-402.
Lyon, R.F., 1996. The all-pole gammatone filter and auditory models, Acustica. Citeseer.
Lyon, R.F., 1997. All-pole models of auditory filtering. Diversity in auditory mechanics, 205-211.
Lyon, R.F., 2011a. Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function. The Journal of the Acoustical Society of America 130, 3893-3904.
Lyon, R.F., 2011b. Using a cascade of asymmetric resonators with fast-acting compression as a cochlear model for machine-hearing applications.
Lyon, R.F., Mead, C.A., 1988. Cochlear hydrodynamics demystified.
Meddis, R., O'Mard, L.P., Lopez-Poveda, E.A., 2001. A computational algorithm for computing nonlinear auditory frequency selectivity. The Journal of the Acoustical Society of America 109, 2852-2861.
Mitra, V., Franco, H., Graciarena, M., Mandal, A., 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4117-4120.
Moore, B.C., 1978. Psychophysical tuning curves measured in simultaneous and forward masking. The Journal of the Acoustical Society of America 63, 524-532.
Moore, B.C., 1985. Frequency selectivity and temporal resolution in normal and hearing-impaired listeners. British Journal of Audiology 19, 189-201.
Moore, B.C., Glasberg, B.R., 1981. Auditory filter shapes derived in simultaneous and forward masking. The Journal of the Acoustical Society of America 70, 1003-1014.
Nelson, D.A., 1991. High-level psychophysical tuning curves: Forward masking in normal-hearing and hearing-impaired listeners. Journal of Speech, Language, and Hearing Research 34, 1233-1249.
Nelson, D.A., Fortune, T.W., 1991. High-level psychophysical tuning curves: simultaneous masking by pure tones and 100-Hz-wide noise bands. Journal of Speech, Language, and Hearing Research 34, 360-373.
Olson, E.S., Duifhuis, H., Steele, C.R., 2012. Von Békésy and cochlear mechanics. Hearing research 293, 31-43.
Patterson, R.D., 1976. Auditory filter shapes derived with noise stimuli. The Journal of the Acoustical Society of America 59, 640-654.
Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function, a meeting of the IOC Speech Group on Auditory Modelling at RSRE.
Qi, J., Wang, D., Jiang, Y., Liu, R., 2013. Auditory features based on gammatone filters for robust speech recognition, 2013 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, pp. 305-308.
Robles, L., Ruggero, M.A., 2001. Mechanics of the mammalian cochlea. Physiological reviews 81, 1305-1352.
Singh, M., Pati, D., 2019. Countermeasures to Replay Attacks: A Review. IETE Technical Review, 1-16.
Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. The Journal of the Acoustical Society of America 106, 2040-2050.
Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., Lee, K.A., 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441.
Villchur, E., 1989. Comments on "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation‐transfer function". The Journal of the Acoustical Society of America 86, 425-427.
Walters, T.C., 2011. Auditory-based processing of communication sounds. University of Cambridge.
Wickramasinghe, B., Ambikairajah, E., Epps, J., 2019a. Biologically Inspired Adaptive-Q Filterbanks for Replay Spoofing Attack Detection, INTERSPEECH, pp. 2953-2957.
Wickramasinghe, B., Ambikairajah, E., Epps, J., Sethu, V., Li, H., 2019b. Auditory Inspired Spatial Differentiation for Replay Spoofing Attack Detection, ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6011-6015.
Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., Gałka, J., 2017. Audio Replay Attack Detection Using High-Frequency Features. Proc. Interspeech, 27-31.
Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015. Spoofing and countermeasures for speaker verification: a survey. Speech Communication 66, 130-153.
Yin, H., Hohmann, V., Nadeu, C., 2011. Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. Speech communication 53, 707-715.