An Adaptive-Q Cochlear Model for Replay Spoofing Detection

Tharshini Gunendradasan[1], *Eliathamby Ambikairajah*[1,2], *Julien Epps*[1,2], *Haizhou Li*[3]

[1]School of Electrical Engineering and Telecommunications, UNSW, Australia
[2]ATP Research Laboratory, DATA61, CSIRO, Australia
[3]National University of Singapore, Singapore
tharshini.gunendradasan@student.unsw.edu.au, e.ambikairajah@unsw.edu.au, j.epps@unsw.edu.au, haizhou.li@nus.edu.sg

Abstract

Replay attack poses a key threat for automatic speaker verification systems. Spoofing detection systems inspired by auditory perception have shown promise to date, however some aspects of auditory processing have not been investigated in this context. In this paper, a transmission line cochlear model that incorporates an active feedback mechanism is proposed for replay attack detection. This model compresses the considerable energy variation in each auditory sub-band filter by boosting low-amplitude signal, an effect that is not considered in many auditory models. To perform the compression, the parameters of each auditory sub-band filter are modified based on the sub-band energy, analogous to the effect of the closed-loop adaptation mechanism that allows perception of a wide dynamic range from a physically constrained system, which we term adaptive-Q. Evaluation on the ASVspoof 2017 version 2 database suggests that the adaptive-Q compression provided by the proposed model helps to improve the performance of replay detection, and a relative reduction in EER of 26% was achieved compared with the best results reported for auditory system-based feature proposed for replay attack detection.

Index Terms: Replay attack, spoofing, speaker verification, Transmission line cochlear model, Active cochlea.

1. Introduction

Automatic speaker verification (ASV) system aims to verify the identity of a person based on her/ his voice. Regardless of the effectiveness of ASV in distinguishing a person's voice from others, it still remains susceptible to spoofing attacks [1] where an unauthorized user uses an illegitimate speech signal that sounds like an original speaker to spoof the ASV system. This spoofing attack is categorized into four types: replay, speech synthesis (SS), voice conversion (VC) and impersonation [2],[3],[4],[5]. Among these attacks, replay is the most easily-performed attack, and doesn't require any expertise, where a recorded voice of a legitimate user is played back in place of genuine speech to manipulate the ASV system. The recent experimental results released on the ASVspoof 2015 and 2017 challenge demonstrate that the replay attack is most difficult to be detected [6], [7].

In the recent years, various front-end features have been proposed for replay detection. It has been shown that high frequency regions are more discriminative to detect replay attack in [8]. Mel filter bank slope and linear filter bank slope features were proposed in [9] to capture the discriminative information present in the low frequency region for low quality devices and the high frequency region for high quality devices, respectively. Further, Constant-Q Cepstral Coefficients (CQCCs) [7] and Linear Frequency Cepstral Coefficients (LFCCs) were proposed and used as the baseline features for replay detection.

Apart from the above features, auditory model-based features were proposed for replay detection. Parallel filter banks that use Gabor and Butterworth filters for bandpass filtering of the signals to extract spectral information in each frequency band was proposed in [10],[11],[12]. In [13], auditory filter banks were learned automatically using Convolutional Restricted Boltzmann Machine (ConvRBM). All these systems extracted amplitude modulation (AM)/ frequency modulation (FM) components from the bandpass signals as a front-end feature.

The above-mentioned auditory models do not really account for much of the auditory system's remarkable ability to analyze complex sounds. In the authors' previous work, the passive transmission line cochlea (TLC) model [14] was proposed, which more accurately resembles the auditory system compared with the above-mentioned parallel filter bank models. The advantage of the sharp roll-off frequency response achieved by the TLC model was discussed in [14], where specifically the amplitude modulation (AM) feature extracted from the TLC model was shown to be very useful for replay detection.

In both the transmission line cochlear model and parallel filter bank model, the input and output signal vary in the same dynamic range. For large energy variation in the input signal, the output signal also changes in the same range. Due to this, it is difficult to capture the discriminative information present in the low energy region, as this energy becomes negligible compared with the high energy present in the signal. In replay attacks, due to reverberation the spectrogram will be temporally stretched, specifically the high energy region will be diffused. It is hypothesized that this kind of small spectral change can be effectively detected by boosting the low energy signal. To perform these kinds of tasks, the cochlea has an active mechanism which compresses the input signal range into a small range by boosting the low energy signal, keeping the high energy signals unchanged. This helps to sense a large dynamic range of signal and capture the information across all energy values.

Various computational active cochlear models have been proposed in the literature. Different variations of active transmission line cochlear models that model the cochlea as a cascade of digital filters were proposed in [15],[16],[17]. In these models, applying compression at a specific section of the membrane will affect other parts and create a masking effect. In this paper, a novel front-end feature for replay detection is proposed. It modifies the passive transmission line cochlear

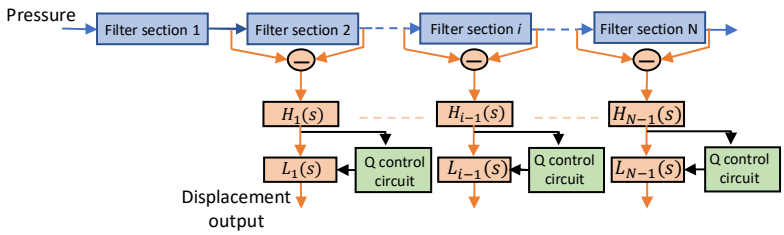


Figure 1. Adaptive-Q cochlear model. Each filter section consists of lowpass, resonant and notch filters. The blue arrows represent the pressure along the basilar membrane and orange arrows represent the process of transforming pressure into displacement. $H(s)$ is the bandpass filter and $L(s)$ is the second-order lowpass filter. Based on the output of $H(s)$, the Q-control circuit updates the Q factor of $L(s)$.

model in [14] to incorporate the active mechanism in the cochlea, which controls the membrane displacement independently without affecting other parts of the membrane.

2. Active transmission line cochlear model

As described in Section 1, this paper proposes modeling the front-end of the spoofing detection as an active cochlear model capable of dynamic range compression, such that the features derived based on this model capture discriminative information present in the low energy region.

2.1. Background: Passive transmission line cochlear model

A one-dimensional computational passive cochlear model was proposed in [18], where each section of the basilar membrane (BM) was modelled using a filter section consisting of lowpass, resonant and notch filters connected in series. The input to each filter section is the pressure $v_i(s)$, and the output pressure $v_o(s)$ is transferred to next filter section. This way the pressure wave will travel along the BM from the high frequency end to the low frequency end. The pressure transfer function of a given filter section is [18]

$$\frac{v_o(s)}{v_i(s)} = K\frac{a}{s+a}\,\frac{{\omega_p}^2}{s^2+b_p s+{\omega_p}^2}\,\frac{s^2+b_z s+{\omega_z}^2}{{\omega_z}^2}, \tag{1}$$

where $0 < K < 1$ is a constant and a is the lowpass filter coefficient, ω_p and ω_z are the resonant and notch angular frequency, respectively and, $Q_p = {\omega_p}/{b_p}$ and $Q_z = {\omega_z}/{b_z}$ are the quality factor (Q-factor) of the resonant and notch filter. The pressure transfer function (Eq. (1)) contains the displacement transfer function as well [18], consequently the displacement of the basilar membrane was obtained from each filter section using only the displacement transfer function (poles only) [18].

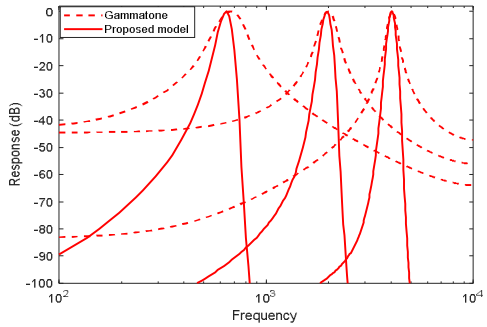


Figure. 2. Frequency response of the proposed model and gammatone filter bank model. Here the Q factors of the gammatone filters were adjusted to obtain similar frequency responses to the proposed model.

2.2. Proposed adaptive-Q cochlear (AQC) model

2.2.1 Proposed modified cochlear model

The block diagram of the proposed modified model is shown in Figure 1. The pressure along the basilar membrane is modelled in the same way as in [18]. The BM displacement is modelled by converting pressure into displacement by adding a lowpass filter ($L(s)$) in parallel as shown in Figure 1. Connecting the lowpass filters in in this manner allows to change the basilar membrane displacement of a particular section independently without affecting other sections. The transfer function of the lowpass filter is

$$L(s) = \frac{{\omega_l}^2}{s^2+b_l s+{\omega_l}^2}, \tag{2}$$

where ω_l is the resonant angular frequency and $Q_l = {\omega_l}/{b_l}$ is the quality factor (Q-factor) of the lowpass filter.

The frequency response of the BM displacement at a particular section can be considered as a bandpass filter, where each section of the basilar membrane resonates at a specific frequency. In the cochlea, due to the coupling effect between adjacent sections, the displacement response roll-off will increase further. To incorporate this coupling effect, spatial differentiation and bandpass filtering are performed. The pressure output at the adjacent filter sections are subtracted as illustrated in Figure 1 to obtain the spatially differentiated signal [14]. For convenience of illustration, only the first-order differentiation is shown in Figure 1. In the proposed model, two orders of differentiation are performed by repeating the same process. Following this, a bandpass filter $H(s)$ is applied,

$$H(s) = \frac{b_b s}{s^2+b_b s+{\omega_b}^2}, \tag{3}$$

where ω_b is the resonant angular frequency and $Q_b = {\omega_b}/{b_b}$ is the Q-factor of the bandpass filter. Then the output signal of the bandpass filter is passed through $L(s)$ to obtain the displacement output.

Figure 2 compares the frequency response of the BM displacement of the proposed model with the gammatone filters, which are the widely used filter banks to represent the auditory system response. Compared with gammatone filters, the proposed model has sharp frequency roll off response, which enables the capture of accurate spectral information within a specific frequency band without capturing any overlapping information from the adjacent frequency bands.

2.2.2 Incorporation of active feedback

The cochlea in the inner ear has an active mechanism which performs dynamic range compression, compressing large-

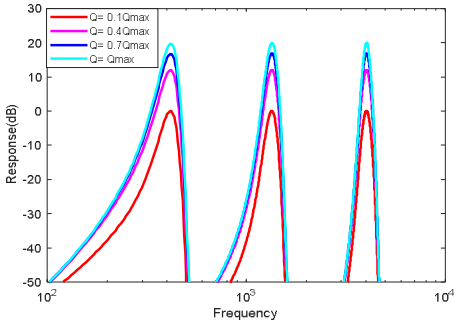


Figure 3. Frequency response of the basilar membrane displacement of active cochlea at low, medium and high frequency when the lowpass filter Q factor is reduced from 100% to 10% of its maximum Q factor value in four equal steps.

amplitude signals into a smaller range. This is done by boosting low-amplitude signals without altering high-amplitude signals.

In the proposed model, the output signal energy is controlled by changing the gain of the lowpass filter $L(s)$. In the lowpass filter $L(s)$, the gain can be directly controlled by changing the quality factor of the filter Q_l, which is equivalent to the gain of the filter at the resonant frequency. During compression, by default lowpass filters are set at their minimum Q value, and for a low energy signal, Q is modified to a higher value, whereas for high energy signals it remains at the low Q value. This way the displacement of the basilar membrane is contained within the physical limits.

Figure 3 shows the frequency response of the basilar membrane displacement when the Q factor of the lowpass filter is varied from the maximum Q factor of the filter (Q_{max}) to the minimum Q factor (Q_{min}) in four steps in a piecewise manner, where $Q_{min} = 0.1Q_{max}$. Changing from minimum Q to maximum Q, the proposed active cochlea model achieved a further 20 dB of gain, which enables it to perform up to 20 dB of compression on the incoming signal.

Figure 1 illustrates the feedback path incorporated into the modified cochlea model. The bandpass filter output is fed into the Q-control circuit and based on the energy of the signal the new Q is estimated and fed into the lowpass filter. For the convenience of illustration, only the bandpass filter output of the particular section is shown for Q estimation. However, the outputs of adjacent filter sections are also taken into account for the robust estimation of the Q factor. The control signal E used to calculate the Q factor at a particular section is the peak output in dB SPL (sound pressure level) over the time interval T found in the range of N adjacent filters in both sides of the

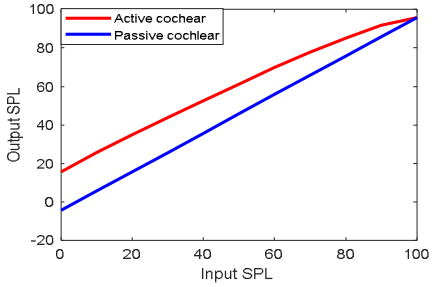


Figure 4. Relationship between the input and output sound pressure levels (SPL) of the active and passive cochlea at the 1kHz tap location for the input stimulus level of 1kHz sinusoidal signal varying from 0 to 100 dB SPL.

filter section. The adaptive Q factor (Q_{adp}) of the lowpass filter is updated every T interval based on E and is varied between maximum (Q_{max}) and minimum (Q_{min}) Q factors of the lowpass filter. Q_{adp} is estimated similar to [17]

$$Q_{adp} = \begin{cases} Q_{min} & E \geq E_{max} \\ Q_{min} + \frac{(Q_{min}-Q_{max})}{(E_{max}-E_{min})}(E-E_{max}) & E_{min} < E < E_{min} \\ Q_{max} & E \leq E_{min} \end{cases}, \quad (4)$$

where E_{max} and E_{min} are the maximum and minimum peak outputs of the bandpass filter. For the input signal ranging from 0 to 100 dB SPL, E_{min} is the peak output of the bandpass filter when the input is 0 SPL sinusoidal signal of frequency corresponding to that filter section. Similarly, E_{max} is the output for the 100 dB SPL sinusoidal signal. In this paper, the proposed active cochlear model is referred to as Adaptive Q-cochlear model (AQC)

Figure 4 shows the input-output relationship of the passive cochlea and the active cochlea models when the input signal level ranges from 0 to 100dB SPL. For the passive cochlear, the output ranges from -4 to 96 dB SPL, the dynamic range of the signal remains the same for both input and output. Whereas the active cochlea compresses the output signal range between 16 to 96 dB SPL, resulting in 1.25 compression ratio.

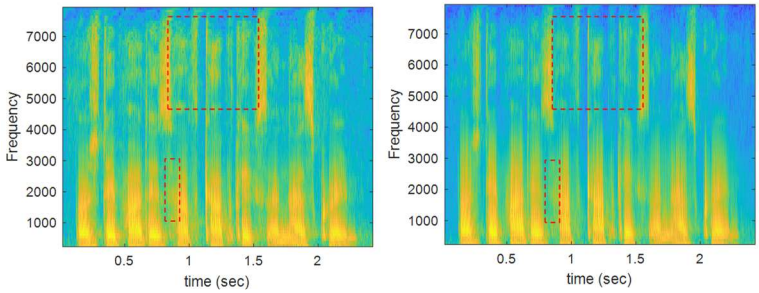
Figure 5 compares the spectrograms obtained using passive and active cochlea. It can be clearly observed that the active cochlear boosts the low energy region, thus in the spectrogram the low energy region becomes more visible for the active cochlear than for the passive cochlear. This enables effective capture of the discriminative information present in the low energy region.

3. Experimental Results

3.1. Design parameters and experimental setup

In the implementation of the AQC model, 110 filter sections were used to represent the cochlea, and the resonant frequency (ω_p) of the resonant filter in each filter sections were placed at equal scales between 100 to 7950 Hz. The notch frequency ω_z was chosen as 1.15*ω_p, Q_p of the resonant filter was varied linearly between 1.2 to 5 from the low frequency end to the high frequency end and Q_z was set to be $2Q_p$, the lowpass filter cutoff frequency was $1.14\omega_z$ and $K = 0.88$ (< 1). The bandpass filter ($H(s)$) and the second order lowpass filter ($L(s)$) were designed to have the same resonant frequency as the corresponding filter section. Q_b was set as 4.5 and kept constant across all the filters. For $L(s)$, the maximum Q-factor was frequency-dependent, determined by the relationship $Q_{l,max} = 1.5(1 + f)$; f is the resonant frequency of the filter section in kHz, and the minimum Q factor was set to $Q_{l,min} = 0.1Q_{l,max}$. The control signal estimation considered eight adjacent filters ($N = 8$) and the Q factors of the second-order low pass filters were updated every 2ms ($T = 2ms$).

For replay detection, amplitude modulation (AM) features were extracted from the AQC model displacement output. AM Feature extraction was performed similarly to [14]. The displacement output of each filter section was full-wave rectified as opposed to half-wave rectified in [14] and then low pass filtered. The low pass signal was averaged over the frame size 2ms, with 50 % overlap. Then log compression and discrete cosine transform (DCT) was performed to extract the features, and the delta and acceleration coefficients were also included along with the static feature. The speech signals were pre-emphasized to emphasize the high frequency region. For



(a) Active cochlea model (b) Passive cochlea model

Figure 5: Active and passive cochlea spectrograms. In the active cochlea spectrogram, low energy regions are more emphasized.

feature-normalization, mean variance normalization (MVN) was performed.

To experiment how the active mechanism contributes to improving the performance of replay detection, the results on the proposed model without the feedback were also generated. The model without feedback loop is referred to as a passive cochlea (PC) model.

For the proposed front-end features, two Gaussian mixture model (GMM) with 512 mixture components were used as the back-end classifier.

The proposed model was tested on the ASVspoof 2017 version 2 database [7]. It contains training, development and evaluation sets of genuine and replayed audio samples. The evaluation set consists of various replayed samples that were obtained under the combination of different replay environments, playback devices, and speakers which are not present in the training and development data in order to evaluate the performance in unforeseen conditions.

3.2. Results and Discussion

This section experiments the proposed AQC model for replay detection and compares it with other state of the art auditory system-based features for replay detection.

Table 1 compares the Equal Error Rate (EER) obtained for AQC and PC models. For PC, the results were reported when the lowpass filter Q factor was at its minimum and maximum values. For AQC, the Q-factor was varied between these two values. At low and high Q values, PC-AM gave nearly the same EERs 8.93 and 9.00, respectively. The EER was reduced to 8.09 for AQC with a relative improvement of 9.4% compared with low Q value PC-AM. The improved performance achieved with this AQC model suggests that boosting the low energy regions helps to improve the replay detection accuracy, which in turn suggests that low energy regions contain discriminative information for replay detection.

As mentioned in *Section 2.2.2*, the proposed method achieved a compression ratio of 1.25, whereas the auditory system can perform high compression, to nearly 2.5. Thus, future work will focus on increasing the compression ratio by modifying the proposed AQC model with the anticipation that emphasizing low energy signal region further and reducing the dynamic range of the signal will help to improve replay detection system performance.

Table 1: Results on ASV spoof 2017 version 2 evaluation data for AQC -AM and PC -AM.

Features	EER
AQC -AM	**8.09**
PC - AM (low-Q)	8.93
PC - AM (High-Q)	9.00

Table 2 compares the performance of the proposed AQC -AM feature with other auditory based models proposed for replay detection in the recent years and the baseline CQCC feature provided by the ASVspoof2017 challenge organizers. Except AQC, the other two auditory models used parallel filter banks as opposed to cascaded filter banks. AQC-AM outperforms other models with the relative improvement of 26.0% and 33.9% compared to the best results reported for parallel filter bank based auditory model CM and the CQCC baseline system, respectively. The significant performance improvement achieved by the proposed model suggests that the AQC model, which more accurately resembles the auditory system compared to the passive parallel filter bank models, taking into account a more realistic high frequency roll-off response and incorporating dynamic range compression, is a good front-end feature extractor for detecting spoofing attacks.

Table 2: Comparison of proposed method with other auditory system-based methods and CQCC feature on ASV spoof 2017 version 2 evaluation data.

Features	EER
AQC -AM	**8.09**
CM [19]	10.93
TECC [20]	11.73
CQCC (baseline) [7]	12.24

4. Conclusions

In this paper, a novel model of the auditory system, the adaptive Q cochlear model (AQC), was proposed, which incorporates the dynamic range compression behavior of the auditory system. It models the auditory system as a cascade of filters. The Q-factor of the second-order lowpass filter connected to the transmission line model was varied with time to compress the dynamic range of the signal. The spectral information obtained showed that the low energy regions are more emphasized and become more visible in the AQC model output than in the passive cochlear (PC) model output. The performance improvement achieved by using the AQC over PC model in ASV spoof 2017 version 2 database was due to the low energy spectral components being boosted, which suggests that discriminative information for replay detection are present in the low energy regions as well. The results obtained show that use of the AQC model leads to a relative improvement of 26% over the best results reported for a parallel filter bank based auditory model. This opens an interesting avenue for the use of AQC models in future front-ends.

5. References

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

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

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication, vol. 51, no. 11, pp. 1039-1064,* 2009.

[4] M. Correia, A. Abad, and I. Trancoso, "Anti-spoofing: Speaker verification vs. voice conversion," *stituto Superior Técnico Master's Thesis,* 2014.

[5] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," *in Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 145-148,* 2004.

[6] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, 'ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,' in *Proc. INTERSPEECH 2015, pp 2037-2041,2015*.

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

[8] 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.

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

[10] 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.

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

[12] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection," *Proc. Interspeech 2017,* pp. 12-16, 2017.

[13] 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.

[14] T. Gunendradasan, S. Irtza, E. Ambikairajah, and J. Epps,'Transmission Line Cochlear Model Based AM-FM Features for Replay Attack Detection,' *Proc. ICASSP 2019, pp. 6136-6140,* 2019.

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

[16] 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.

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

[18] 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.

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

[20] M.R. Kamble, and H.A. Patil, 'Analysis of Reverberation via Teager Energy Features for Replay Spoof Speech Detection,' *Proc. ICASSP 2019, pp. 2607-2611*, 2019.