Voice Liveness Detection using Constant-Q Transform-Based Features

Ankur T. Patil, Kuldeep Khoria, Hemant A. Patil
Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India
Email: {ankur_patil, kuldeep_khoria, hemant_patil}@daiict.ac.in

***Abstract*—In this work, we propose to use the Constant-Q transform (CQT)-based feature set for voice liveness detection (VLD), which can enhance the confidence in authenticity of the speaker in Automatic Speaker Verification (ASV) system. The live speaker can be characterized via his/her voice using the presence of the pop noise in the speech signal. Pop noise comes out as a burst and possesses the low frequency characteristics. In this paper, we present the modified CQT-based approach over the traditional Short-Time Fourier Transform (STFT)-based algorithm (baseline) for VLD. The experiments are performed on recently released *POp noise COrpus* (POCO) dataset with various statistical, discriminative, and deep learning-based classifiers, namely, Gaussian Mixture Models (GMMs), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Light-CNN (LCNN), respectively. The significant improvement in performance is observed for the proposed CQT-based features over STFT-based features. Relatively best performance is obtained for CQT-LCNN architecture, which shows 81.93% accuracy on evaluation set. Furthermore, we analyzed the performance of the CNN and LCNN-based VLD systems for each word using proposed CQT-based *vs.* STFT-based baseline features.**

***Index Terms*—Constant-Q transform, STFT, Voice liveness detection, Pop noise, POCO dataset.**

I. Introduction

The use of Automatic Speaker Verification (ASV) for authentication is increasing in many applications, such as voice assistant, financial transactions, and various applications in smartphones [1]. However, ASV is susceptible to various spoofing attacks, namely, speech synthesis (SS), voice conversion (VC), mimicry, twins, and replay attacks [1], [2]. Among these attacks, replay attack is easier to mount but difficult to detect because of availability of high quality microphones and loudspeakers. This work presents the novel countermeasure (CM) system or spoof speech detection (SSD) system by exploiting the characteristics of the pop noise event in the live (genuine) speech signal.

Various feature sets and classifiers are proposed to build the CM systems against various spoofing attacks using the publicly available datasets, released during ASVSpoof challenge campaigns [3]. However, less work is reported in the direction of voice liveness detection (VLD) of the speaker to validate his/her authenticity. To the best of authors' knowledge, the problem of VLD was introduced for the first time in [4], where possible methodologies of the pop noise (liveness) detection were discussed. Phoneme-based pop noise detection is performed in [5], where pop noise duration is detected in the utterance and estimated phonemes in this duration are analyzed for liveness detection. This approach is further extended with Gammatone Frequency Cepstral Coefficients (GFCC) feature set for pop noise detection [6].

This paper is an extension of our initial work in [7], where constant-Q transform (CQT)-based feature set was proposed for VLD using recently released POp noise COrpus (POCO) dataset [8]. In [7], the performance of the CQT was analyzed using support vector machine (SVM) classifier, whereas, in this work, we present more detailed analysis of CQT *vs.* Short-Time Fourier Transform (STFT) and experiments are extended using Gaussian Mixture Model (GMM), Convolutional Neural Networks (CNN), and Light-CNN (LCNN). The key motivation of using CQT for VLD is its high frequency resolution in low frequency regions and hence, CQT is capable of capturing the prominent acoustic cues for pop noise, present in the low frequency regions. The evolution of the CQT can be sequentially studied in [9]–[15]. As original investigations in Youngberg's work, we utilized Hanning window as an analysis window in this work for both CQT and STFT [16]. The rest of the paper is organized as follows. Section II describes the details of the proposed CQT-based feature set. Section III and Section IV gives the details of the experimental setup and results, respectively, for this work. Finally, Section V concludes the paper along with future research directions. The mathematical structure of CQT, proposed CQT-based features, experimental setup, and results are discussed in subsequent sections.

II. Proposed Approach

In [4], the features for pop noise detection, are derived from STFT. The same algorithm is used in [8] for VLD on POCO dataset. Therefore, we considered it as a baseline approach, where spectral energy densities of the speech signal are estimated using STFT spectrogram. DFT is nothing but uniform sampled version of DTFT performed on each frame of the speech signal [15]. Let $x(n)$ be the discrete-time input speech signal having sampling frequency F_s. Then, STFT of $x(n)$ is given by [17]:

$$X(\omega, \tau) = \sum_{n=-\infty}^{\infty} x(n) \cdot w(n, \tau) \cdot e^{-j\omega n}, \qquad (1)$$

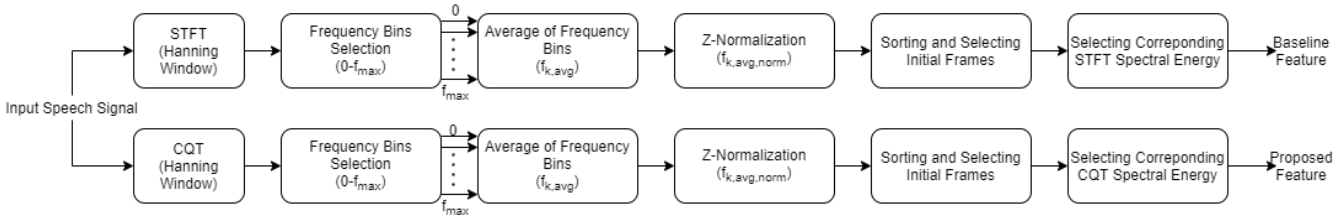

Fig. 1: Block diagram of baseline and proposed algorithm. After [7].

TABLE I: Window length in samples as a function of analysis frequency (f_k). After [15].

k	Frequency (Hz)	# Samples	Duration (in s)
1	0.6729	4390912	199.13
100	1.3753	2148411	97.43
400	12	246262	11.16
800	215	13712	0.62
1345	11025	268	0.0122

where $w(n,\tau)$ represents the analysis window, centered at time τ. It should be noted that $w(n,\tau)$ is function of *only* time parameter τ as independent variable. Furthermore, let $y(n)$ represents a frame of the speech signal, then the DFT, $Y(k)$, of the $y(n)$ can be represented as [18]:

$$Y(k) = \sum_{n=0}^{N-1} y(n) \cdot e^{-j(\frac{2\pi}{N})kn}, \tag{2}$$

where k is the frequency bin index, $\omega_{DFT} = (2\pi k)/N$. In the proposed algorithm, we have employed CQT instead of STFT in order to obtain the high resolution frequency bins in low frequency regions. In CQT, the quality factor Q of the subband filters in the filterbank remains constant and hence, frequency bins are geometrically-spaced. The CQT of a signal $y(n)$ is represented as:

$$Y^{CQT}(k) = \frac{1}{N(k)} \sum_{k=0}^{N(k)-1} y(n) w(n,k) e^{-j\left(\frac{2\pi}{N(k)}Qn\right)}, \tag{3}$$

where $\omega_{CQT} = (2\pi Qn)/N(k)$ and analysis window $w(n,k)$ is *Hanning* window in this paper which has the identical shape for analysis of each frequency component f_k, however, its length is determined by $N(k)$ and thus, it is function of both n and k, where $N(k) = Q(F_s/f_k)$. Table I shows the window length for the set of parameters of CQT for our application. It can be observed from Table I that the window length varies w.r.t. f_k, and it reduces with increase in f_k. Window length is very high for lower frequency regions, which provides the high frequency resolution and hence, the pop noise characteristics in low frequency regions can be effectively captured by the CQT. Here, the quality factor Q is the ratio of center frequency to the bandwidth of each window and it is given by [15]:

$$\because Q = \frac{f_k}{\Delta f_k} = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/B} - 1}, \tag{4}$$

where B represents the number of bins per octave, and f_k represents the frequency of k^{th} spectral component which is given as $f_k = (2^{(k-1)/B}) f_{min}$, where f_{min} is the minimum frequency of the signal. Furthermore, we resampled the magnitude spectrum of CQT to linear scale in order to reduce the number of frequency bins in the feature set [19].

To extract the features for pop noise detection, we employed the similar processing on resampled CQT as it was described for STFT in [4], [8]. Let S_{eng} be the spectral energy densities of the initial frequency bins which corresponds to 0-f_{max} Hz. Then, $f_{k,avg}$ is estimated as average of the spectral energy densities of the STFT spectrogram (where k is the frame index) by applying averaging operation across the bins on S_{eng} for each frame. Then, mean and standard deviation is estimated for averaged spectral energies $f_{k,avg}$. Now, this mean and standard deviation is used for normalization of $f_{k,avg}$ to obtain $f_{k,avg,norm}$. Then, 10 frames were chosen with the largest spectral energies. This is done by taking 10 frames from $f_{k,avg,norm}$ having largest values and then taking frames corresponding to that indices from S_{eng}. For STFT, f_{max} was chosen as 40 Hz. Empirically, we observed that resampled CQT produces relatively better performance for f_{max} = 30 Hz and hence, we utilized this value of f_{max} in this work. Other parameters of CQT chosen for our application are as follows: f_{min} = 0.67 Hz, B = 96, and Q = 134 (opposed to Q = 34 for western music notes [15]). Table II shows the comparison of STFT *vs.* CQT for various spectral parameters utilized in this study. It can be observed that resampled CQT requires less number of frequency bins as compared to the original CQT for representing the f_{40Hz}. For fair comparison, we also performed the experiment for STFT using 120 number of frequency bins to represent the 0 to f_{max} Hz, which uses the frequency resolution of 0.33 Hz. However, we observed the better performance for the frequency resolution of 1 Hz and hence, further experiments in this work are performed with this value of frequency resolution.

III. Experimental Setup

A. Dataset Used

In this work, the recently released POCO dataset is used. The details of the dataset can be studied from [8]. Among the three subsets in POCO dataset, *RC-A* (genuine) and *RP-A* (spoof) subsets are utilized for the experiments. The utterances in *RC-A* subset consists of pop noise and hence, it is considered as genuine utterance as it represents the liveness. *RP-A* subset does not consists of pop noise and hence, it can be treated as spoof utterance. The *RC-A* and *RP-A* subsets are

TABLE II: The comparison of the CQT, *resampled* CQT, and STFT w.r.t. the various spectrographic parameters for pop noise detection with $F_s = 22050\ Hz$. After [15].

Parameters	CQT [15]	Resampled CQT [19]	STFT
f_{min}	0.67 Hz	0.67 Hz	1 Hz
$f_{Nyquist}$	11050 Hz	11050 Hz	11050 Hz
f_{40Hz}	40 Hz	40 Hz	40 Hz
# bins for $f_{Nyquist}$	1345	32850	11050
# bins for f_{40Hz}	567	118	40
Resolution	varying = $\frac{f \cdot k}{Q}$	0.3365 Hz	1 Hz
Q	Constant	Variable	Variable

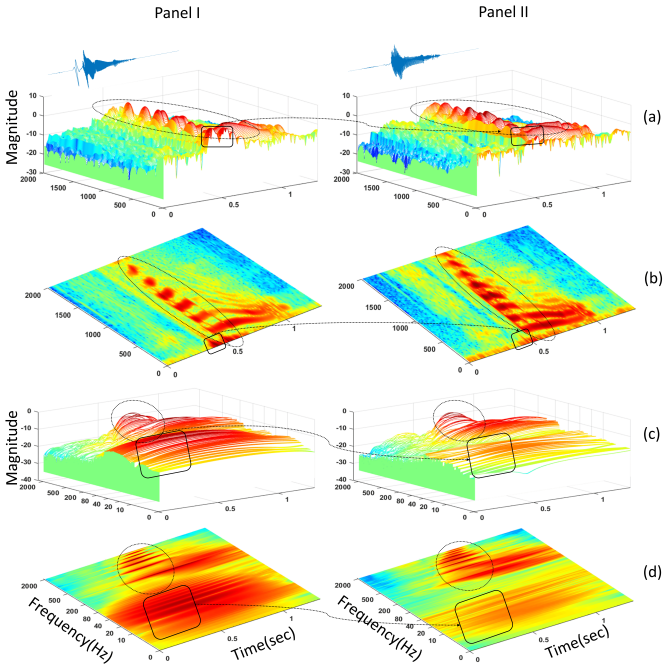

Fig. 2: Panel-I and Panel-II depicts the spectrographic analysis for word "funny" for genuine and spoof speech signal, respectively. (a) the waterfall plot for STFT, (b) the top-view of the STFT waterfall plot, (c) waterfall plot for CQT, and (d) the top-view of the CQT waterfall plot.

further partitioned into training, development, and evaluation subsets as 40%, 20%, and 40% utterances, respectively. The detailed statistics of this partition is shown in Table III.

B. Classifiers and Performance Metrics

Four classifiers are employed in this work, namely, GMM, SVM, CNN, and LCNN. The SVM was utilized in [8] and hence, we employed it as a classifier for baseline architecture. SVM is a non-probabilistic binary linear classifier which gives an optimal hyperplane for given labeled training data, and categorizes new examples [20]. A GMM is represented by a weighted sum of the individual pdfs parameterized by a number of mean vectors, covariance matrices, and mixture weights [20]. We have performed experiments by varying the number of Gaussian mixture and empirically we have observed that 512 number of Gaussian mixture gives us the optimum results.

Furthermore, two deep learning-based classifiers are employed, namely, CNN and LCNN. The CNN used in this work consists of four convolution layers and 1 Fully-Connected (FC) layer. The output of these four convolutional layers have 4, 16, 32, and 8 channels, respectively, which are followed by max-pooling layers. Rectified Linear Unit (ReLU) function is used as the activation function in the hidden layers. The model is trained using Stochastic Gradient Descent (SGD) algorithm with a batch size of 64, and learning rate of 0.001. Binary cross-entropy loss is chosen as the loss function. The experiments are executed for a total number of 400 epochs. The LCNN architecture was employed here since it is one of the successful architectures for replay SSD task [21], [22]. LCNN architecture uses Max-Feature-Map (MFM) activation operation, which is a special case of max-out, for learning with a small number of parameters [23]. In comparison with ReLU, for which the operating threshold is learned from the training data, MFM uses an interesting strategy exhibiting a better generalization ability for distinct data distributions.

In this work, two performance metrics, namely, Percentage Accuracy and Equal Error Rate (EER) are used for measuring the performance of the proposed systems and to evaluate their classification performance [24].

TABLE III: Statistics of the POCO dataset used for our experiments. After [8].

Subset	# Utterances	# Speaker	# Male	# Female
Training	6952	27	13	14
Development	3432	13	6	7
Evaluation	6600	26	13	13

IV. Experimental Results

A. Spectrographic Analysis

Spectral details of the pop noise lie in the low frequency regions. The STFT possesses the constant separation (resolution) between the frequency bins. However, the CQT displays the frequency-domain representation with high frequency resolution at lower frequency regions and vice-versa. Hence, CQT efficiently captures the spectral details of the pop noise. This can be observed from the waterfall plot for word "funny" and it's top view of the STFT- and CQT-gram for the genuine *vs.* spoof speech signal as shown in Fig. 2. The rectangular box in Fig. 2 represents the intended portion of the pop noise, whereas encircled area represents the fundamental frequency ($F0$) and its harmonics. It can be observed that the CQT gives more emphasis on pop noise region than the $F0$ and its harmonics, as compared to its STFT counterpart. The higher resolution property of the CQT allows the pop noise to occupy more area in the CQT-gram with higher intensity as compared to that of STFT-gram. Because of having the larger region for pop noise in CQT-gram, it is much more easier (than its STFT counterpart) for the back-end classifier to discriminate the live (genuine) *vs.* spoof speech signal.

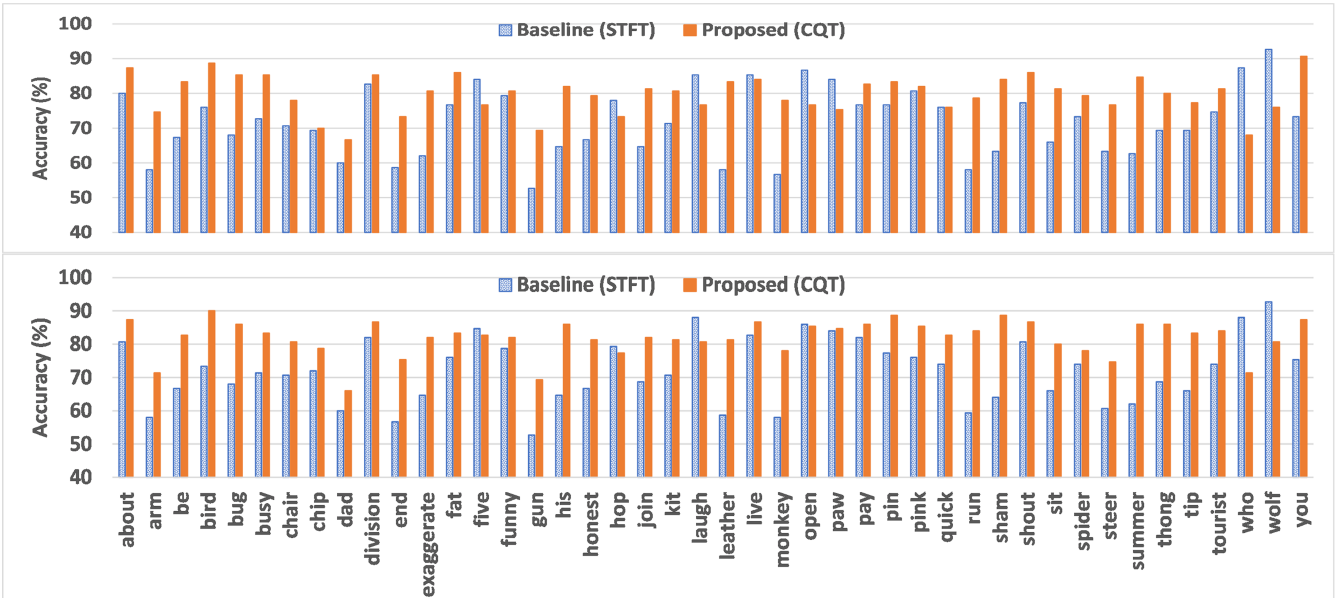

Fig. 3: Comparison of wordwise percentage accuracy on evaluation set for STFT- (baseline) and CQT-based (proposed) feature set with (a) CNN, and (b) LCNN as classifier.

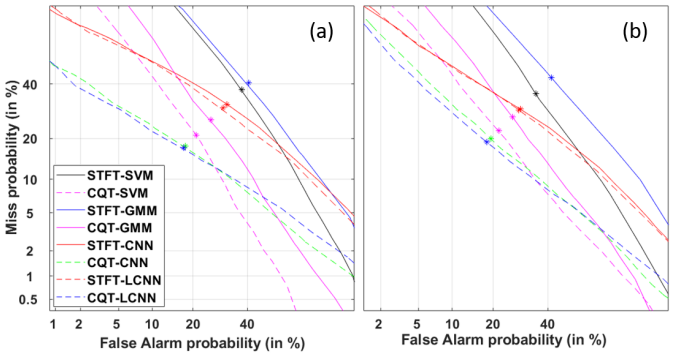

Fig. 4: DET curves for the proposed CQT *vs.* STFT-based baseline algorithm for various classifiers on (a) development, and (b) evaluation set. Legends in Fig. 4(b) are the same as that of Fig. 4(a).

B. Results on Evaluation Metrics

The results in % classification accuracy and % EER for STFT-based baseline features *vs.* proposed CQT-based features using various classifiers, are shown in Table IV. Hanning window is used for both CQT- and STFT-based features. It can be clearly observed that the proposed CQT-based feature set outperforms over the STFT-based baseline features for all the classifiers. Relatively best performance is observed for the proposed CQT-based features using LCNN classifier, which shows 81.93% classification accuracy (i.e., 18.26 % EER) on evaluation set. The similar trends in the performance can be observed in the DET curves obtained using various systems and shown in Fig. 4.

Furthermore, the experiments were performed by increasing the amount of training data in order to majorly investigate

TABLE IV: Results of STFT *vs.* CQT-based feature sets for various classifiers. Hanning window is used for both CQT- and STFT-based features.

Feature Set	Classifier	% Accuracy		EER	
		Dev	Eval	Dev	Eval
STFT	SVM	65.61	67.93	37.61	35.11
STFT	GMM	55.22	53.85	40.42	41.60
STFT	CNN	70.57	71.81	31.80	29.15
STFT	LCNN	70.60	71.90	30.37	28.69
CQT	SVM	79.34	78.42	21.03	21.72
CQT	GMM	73.48	72.59	26.02	26.33
CQT	CNN	82.27	79.77	17.87	19.28
CQT	LCNN	**83.68**	**81.93**	**17.29**	**18.26**

TABLE V: Comparison of proposed approach *vs.* the baseline approach with larger training data (80 % training, 20 % testing) for various classifiers on POCO dataset. Hanning window is used for both CQT- and STFT-based features.

Feature Set	Classifier	Accuracy (%)	EER (%)
STFT	SVM	66.10	38.10
STFT	GMM	54.91	43.50
STFT	CNN	69.90	33.33
STFT	LCNN	71.70	29.52
CQT	SVM	80.40	20.46
CQT	GMM	73.59	26.93
CQT	CNN	84.71	15.87
CQT	LCNN	85.50	15.53

the improvement in the performance for CNN, and LCNN architectures since deep learning architectures are known to perform well for large amount of training data. To that effect, we divided the dataset into two parts, i.e., training and testing with a ratio of 80 % and 20 %, respectively. The corresponding results are presented in Table V. It can be observed that the performance of both the baseline and proposed approaches

with SVM as a classifier is almost the same as stated in Table IV, where dataset division is 40 % training, 20 % development, and 40 % evaluation. However, proposed CQT-based algorithm shows the marginal improvement over all the algorithms.

Furthermore, we computed the wordwise accuracy for the proposed CQT- *vs.* STFT-based features using CNN and LCNN classifiers on evaluation set as shown in Fig. 3. POCO dataset paper mentions the words and the corresponding International Phonetic Alphabet (IPA) recorded for creating the POCO dataset [8]. Given this relation between words and IPA for the phonemes, we could compute the wordwise accuracy. It can be clearly observed from Fig. 3 that around 85% and 90% classification accuracy is obtained (for words, such as 'busy', 'division', 'fat', 'funny', 'five', 'thong', and 'shout') for the proposed CQT-based algorithm using CNN and LCNN classifiers, respectively. In addition, for most of the words, proposed CQT-based features performs relatively better as compared to the STFT-based features.

V. Summary and conclusions

In this study, we exploited the CQT-based features to detect the liveness of the speaker by using the pop noise as a discriminative acoustic cue. The experiments are carried out on recently released POCO dataset. The results of proposed approach are compared against the baseline, where feature set is derived from the traditional STFT. The spectrographic analysis for genuine (live) *vs.* spoof speech is performed which showed that the pop noise is emphasized in a much better way by CQT-based spectrogram than it's STFT counterpart. Furthermore, the experimental results obtained using various classifiers, namely, GMM, SVM, CNN, and LCNN shows the efficacy of the proposed CQT-based features over the traditional STFT-based features, for pop noise detection.

VI. Acknowledgement

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation Mission (NLTM): BHASHINI', subtitled 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy, Prof. S. Umesh, and the authorities of DA-IICT Gandhinagar, India for their support and cooperation to carry out this research work.

References

[1] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre, and Phillip De Leon, "Speaker recognition anti-spoofing," in *Handbook of Biometric Anti-spoofing*, pp. 125–146. Springer, 2014.

[2] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] *ASVspoof challenge campaigns and workshops*, URL: https://www.asvspoof.org/ {Last accessed Oct. 2, 2021}.

[4] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH,Dresden,Germany*, 2015, pp. 239–243.

[5] Shihono Mochizuki, Sayaka Shiota, and Hitoshi Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verification," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, pp. 233–239.

[6] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019, pp. 2062–2070.

[7] Kuldeep Khoria, Ankur T. Patil, and Hemant A. Patil, "Significance of constant-Q transform for voice liveness detection," in *European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, August 2021.

[8] Kosuke Akimoto, Seng Pei Liew, Sakiko Mishima, Ryo Mizushima, and Kong Aik Lee, "POCO: A voice spoofing and liveness detection corpus based on pop noise," in *INTERSPEECH*, Shanghai, China, October 2020, pp. 1081–1085.

[9] Nobert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, MIT Press, Mass, 1949.

[10] Dennis Gabor, "Theory of Communication. Part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[11] Manfred R. Schroeder and Bishnu S. Atal, "Generalized short-time power spectra and autocorrelation functions," *The Journal of the Acoustical Society of America (JASA)*, vol. 34, no. 11, pp. 1679–1683, 1962.

[12] G. Gambardella, "A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters," *IEEE Transactions on Circuit Theory*, vol. 18, no. 4, pp. 455–460, 1971.

[13] Richard A. Altes, "The Fourier–Mellin transform and mammalian hearing," *The Journal of the Acoustical Society of America (JASA)*, vol. 63, no. 1, pp. 174–183, 1978.

[14] G. Gambardella, "The Mellin transforms and constant-Q spectral analysis," *The Journal of the Acoustical Society of America (JASA)*, vol. 66, no. 3, pp. 913–915, 1979.

[15] Judith C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.

[16] James Youngberg and Steven Boll, "Constant-Q signal analysis and synthesis," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tulsa, Oklahoma, USA, 1978, vol. 3, pp. 375–378.

[17] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1^{st} edition, Pearson Education India, 2015.

[18] Alan V Oppenheim, Alan S Willsky, Syed Hamid Nawab, Gloria Mata Hernández, et al., *Signals & systems*, Pearson Educación, 1997.

[19] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "Constant-Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[20] Christopher M Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[21] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 82–86.

[22] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 1033–1037.

[23] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[24] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, 1997, pp. 1895–1898.