Significance of Constant-Q Transform for Voice Liveness Detection

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

***Abstract*—In this paper, we present the novel approach of the liveness detection in speech signal based on constant-Q transform (CQT) which employs geometrically distributed frequency bins. Pop noise can be attributed to the liveness in the speech signal and we exploited this attribute for liveness detection. Pop noise is created due to spontaneous breathing while uttering the certain phonemes which includes the plosive burst, and it has low frequency characteristics. We follow the approach of liveness detection in original POCO dataset paper as baseline, where features are derived from Short-Time Fourier Transform. In our approach, we exploited the low frequency characteristics of pop noise using CQT which has variable spectro-temporal resolution with high resolution at low frequency regions. The experiments are performed on recently released publicly available *POp noise COrpus* (POCO) dataset. The 10-fold cross-validation performed using proposed approach shows improvement in absolute accuracy by 4.2% as compared to the baseline system. The proposed approach also shows relatively better performance for our *disjoint* partition (in terms of speakers) of the dataset.**

***Index Terms*—Voice liveliness detection, pop noise, CQT, SVM, POCO dataset.**

I. Introduction

As voice is the most convenient and natural way of communication, voice biometrics has emerged in many real-time applications, such as financial transactions, commanding the personal devices. This leads to the development of the robust ASV systems. However, this robustness in ASV system made it more susceptible to the spoofing attacks [1]–[3]. Spoofing may be performed through speech synthesis (SS), voice conversion (VC), mimicry, and replay attacks [4]–[7]. To alleviate these issues, the research community in the field of ASV and anti-spoofing has organized several challenges, such as ASVSpoof challenge campaigns during INTERSPEECH conferences [8]–[10]. Many algorithms are developed in these challenges itself and later based on the standard datasets and evaluation metrics provided by these challenge organizers. These algorithms are mainly developed using the spectrogram-based feature sets followed by conventional (GMM, SVM) or deep learning-based architectures (Light-CNN, Siamese Network). Few best performing architectures can be studied in [11]–[17]. Comprehensive review of the challenge campaign can be studied from [18].

These challenge campaigns include the datasets to develop the countermeasure system for VC, SS or replay spoofing attacks that use the distortions introduced by spoofing mechanism as a signature to detect the spoofing attack. However, less attention is given towards the liveness detection to avoid the spoofing attacks. To that effect, recently POCO (POp noise COrpus) dataset is constructed which can be used to build the countermeasure strategies against spoofing attacks by identifying the presence of pop noise present in live, i.e., genuine speaker's voice [19]. Pop noise causes the distortion in the speech signal introduced by the speaker's breath. Thus, it is the characteristic of the live speech. Identifying the pop noise for live speaker detection might be very useful strategy in the applications where the testing microphone is placed at a short distance from the speaker, and consequently this strategy may protect the ASV system from spoofing attacks.

Liveness detection for spoofing detection is proposed for the first time in [20], where two approaches of liveness detection are proposed: (a) low-frequency-based single channel detection, (b) subtraction-based pop noise detection with two channels. In the former approach, Short-Time Fourier Transform (STFT) around lower frequency region is utilized as the pop noise exists in the lower frequency regions. Whereas in the later approach, entire frequency range of the spectrum is utilized. In [21], phoneme-based pop noise detection is performed for liveness detection along with speaker verification system, where pop noise duration is detected in the utterance and estimated phonemes in this duration are analyzed for liveness detection. The similar approach of phoneme-based pop noise detection was utilized in [22] with extended study on Gammatone Frequency Cepstral Coefficients (GFCC) feature set for pop noise detection.

In this paper, we propose to exploit geometric frequency spacing of the constant-Q transform (CQT)-based spectral representation for the liveness detection. In this work, we have replaced the STFT with CQT in the algorithm proposed in [20]. The key motivation of using CQT is its high frequency resolution in low frequency regions by which CQT is capable of capturing the prominent cues for liveness detection present in the low frequency regions. The experiments are performed using 10-fold cross-validation and have obtained the promising results which are discussed in Section IV.

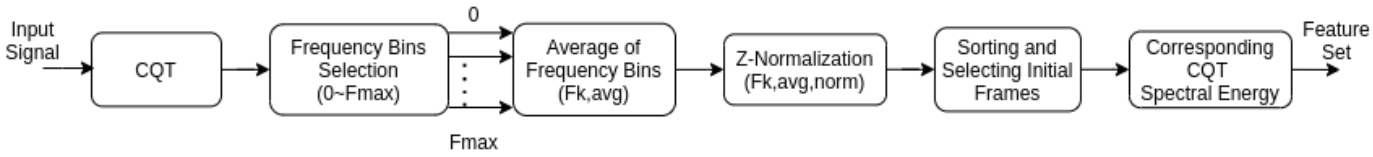

Fig. 1. Block diagram of proposed algorithm

II. Proposed approach

A. Pop Noise

During speech production, airflow travels from the lungs to vocal folds, excites vocal tract system and finally, it bursts out from the mouth as a sound wave. While capturing this sound via microphone, if the distance between speaker and microphone is less, the microphone in addition to capturing speech signal, it also captures the friction between lips and the airflow as *plosive burst* which is termed as *pop noise*. On the other hand, the attacker who is trying to record the voice usually cannot put recording device near to the live speaker and hence, due to increase in distance between live speaker and recording device, pop noise will not be present or with very low intensity when the attacker will replay it to the microphone. Hence, by recognizing the pop noise, we can distinguish between the live (genuine) speech and the replayed speech [23]. Hence, pop noise can act as a important acoustic cue for voice liveness detection systems.

B. Baseline Algorithm

In [20], the features for pop noise detection, are derived from STFT. The same algorithm is used in [19] for liveness detection on POCO dataset. Therefore, we considered it as a baseline approach.

In baseline approach, spectral energy densities of the speech signal is estimated using STFT spectrogram. Let, S_{eng} be the spectral energy densities of the initial frequency bins which corresponds to 0-F_{max} Hz. For this work, F_{max} was chosen as $40Hz$ as pop noise is present at low frequency regions [20]. Then, $F_{k,avg}$ is calculated 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 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}. Utterances from the **RC-A** subset were labeled as positive, and samples from the **RP-A** subset were labeled as negative.

C. Proposed Algorithm

In the proposed algorithm, we have employed constant Q transform (CQT) instead of STFT in order to obtain the high resolution frequency bins in low frequency regions. The block diagram of proposed algorithm is shown in Fig. 1. The

Algorithm 1 Pseudo Code of Proposed Algorithm

1: $F_{cqt} = cqt(x)$, Applying CQT to speech signal,
2: $S_{eng} = (abs(F_{cqt}(1 : F_{bins(40Hz)}, :)))^2$, Taking bins upto $40Hz$ only,
3: **for** *i=1:length*(S_{eng}) **do**
    $F_{k,avg} = mean(S_{eng}(:, i))$, Taking average of CQT spectrogram along column vector,
4: $MN = mean(F_{k,avg})$, $SD = std(F_{k,avg})$, Estimate mean and standard deviation,
5: **for** *i=1:length*($F_{k,avg}$) **do**
    $F_{k,avg,norm} = (F_{k,avg}(i) - MN)/SD$, Normalising,
6: $[F_{k,avg,norm,sort}, index] = sort(F_{k,avg,norm})$, Sorting,
7: $F_{k,avg,intial} = F_{k,avg,norm,sort}(1 : 10, :)$, Taking initial 10 frames,
8: $index_{initial} = index(1 : 10, :)$, Taking corresponding indices,
9: **for** *i=1:length*($F_{k,avg,intial}$) **do**
    $feat = S_{eng}(:, index_{initial}(i))$, Feature set

motivation behind using CQT is that in realistic scenarios of speech production and perception, the frequency of the speech signal doesn't have constant frequency interval rather it has geometrical distribution [24]. As while applying STFT, the subband filters have constant frequency interval, they might be unable to map the frequency content of the speech signal accurately. Therefore, we propose to use CQT instead of STFT as it uses constant-Q ratio of center frequency to resolution and hence, giving the geometrically-spaced subband filters. The center frequency f_c of c^{th} frequency bin is obtained by [24]:

$$f_c = f_l \cdot 2^{\frac{(c-1)}{B}}, \tag{1}$$

where f_l is the center frequency of the lowest frequency bin, and B is the number of bins per octave. The Q factor is given by [24]:

$$Q = \frac{f_c}{f_{c+1} - f_c} = \frac{1}{2^{1/B} - 1}. \tag{2}$$

The pesudo code of the proposed algorithm is illustrated in Algorithm 1. Here F_{cqt} is obtained by uniform resampling of log power magnitude spectrum obtained by applying CQT. Further, the average of CQT spectrogram $F_{k,avg}$ within the interval [0, F_{max}] is computed for each column vector. As pop noise is present at low frequency region, we have considered F_{max} as 40 Hz. In the proposed algorithm, S_{eng} for CQT is computed as it was computed in baseline algorithm for STFT.

Then normalization of $F_{k,avg}$ is done to zero-mean and unit standard deviation to obtain $F_{k,avg,norm}$. Then 10 frames from

TABLE I
STATISTICS OF THE POCO DATASET FOR OUR EXPERIMENTS.

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

$F_{k,avg,norm}$ is considered which have largest values, and then taking frames corresponding to that indices from S_{eng}.

III. EXPERIMENTAL SETUP

If an attacker wants to do a spoof attack, the attacker must somehow obtain the samples of the genuine speaker. The simplest way to do this is by recording (eavesdropping) the voice of genuine speaker and then replaying it in front of ASV system. In realistic scenarios, this recording will be done from long distance due to which the pop noise will not be recorded by the attacker's microphone and there will be absence of pop noise in the replayed sample.

A. Database Used

In this work, we have used recently released POCO dataset [19]. The dataset is sampled at 22050 Hz with a bit rate of 24-bits. There are total of 66 speakers in which 34 are male and 32 are female. The words were selected such that all 44 phonemes in English language were covered in the recording. The dataset has three subsets, namely, RC-A (Recording with Microphone), RP-A (Eavesdropping), and RC-B (Recording with Microphone Array) . We have excluded the RC-B subset for our experiments as it consists of microphone array and it's corresponding spoof speech utterances are not provided. Also the experiments in [19] are performed using RC-A and RP-A subsets. The details of RC-A and RP-A are as follows:

1) Recording with Microphone (RC-A): The recording was done with Audio-Technica AT4040 microphone. This subset represents genuine speaker as it was recorded directly with the live speaker and hence, contains pop noise. The distance between speaker and microphone was fixed to be 10 cm.

2) Eavesdropping (RP-A): Here the scenario is considered where replay attack is done by an attacker and for that recording is done from a long distance, i.e., without pop noise. This is done by using Audio-Technica AT4040 microphone with a pop filter inserted between speaker and microphone. The distance between speaker and microphone was fixed as 10 cm.

The dataset is partitioned into training, development, and evaluation subsets as 40%, 20%, and 40% utterances, respectively. Each of these subsets consist of half of the genuine and half of the spoof speech utterances. We also ensured that the speakers are exclusive in each subset and the ratio between male and female speaker is maintained. The statistics of the data distribution in training, development, and evaluation subset is shown in Table I.

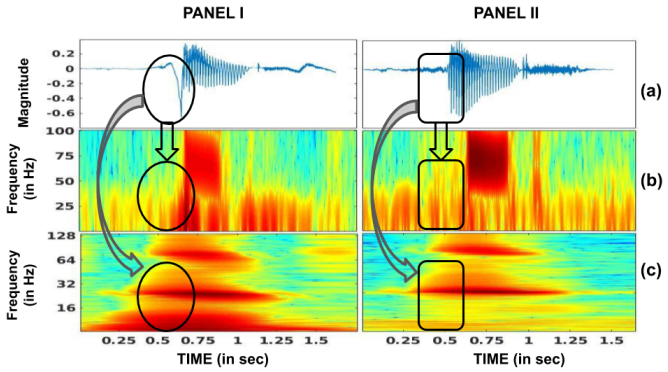

Fig. 2. Spectrogram is shown for word 'thong'. Panel I (a), (b), and (c) shows the speech signal, spectrogram obtained by applying STFT, and CQT respectively for the genuine utterance. Panel II (a), (b), and (c) shows the speech signal, spectrogram obtained by applying STFT, and CQT, respectively, for the spoofed utterance. The pop and non pop locations are highlighted by circle and rectangle box, respectively, for both the spectrograms.

B. Feature Sets, Classifier, and Performance metric

In this study, we used the algorithm explained in Section II-B as a baseline algorithm which is based on STFT. We propose to use CQT-based spectral representation instead of STFT. The CQT seems to be more apt choice for this application as CQT exhibits high frequency resolution for the low frequency regions which is important to analyze the presence of pop noise. We have taken number of bins per octave (B) as 96 to get CQT based spectrum. Number of samples in first octave was set as 2. The minimum and the maximum frequecy for CQT computation is set as $f_{min} = \frac{f_s}{2^{15}}$ = 0.48 Hz and $f_{max} = \frac{f_s}{2}$ = 11025 Hz. We have set f_{min} such that we can capture low frequency regions with high resolution with the help of CQT.

We have selected the Support Vector Machine (SVM) as a classifier as it was used in the baseline [19]. SVM is a non-probabilistic binary linear classifier as it assigns any new data point directly to the one of the classes. The SVM gives an optimal hyperplane given labeled training data which categorizes new examples [25], [26]. Usually, kernel trick is used for transformation of data into suitable form for the classification [25]. We have used 2-class linear kernel for the classification task [26]. Further, L2 regularization is used along with hinge loss for "maximum-margin" classification.

We have reported all the results in terms of % accuracy.

IV. EXPERIMENTAL RESULTS

A. Spectrographic analysis

Fig. 2 shows speech signals and corresponding spectrograms for the genuine and spoofed speech utterances, respectively, after applying baseline and proposed algorithm. Since the pop noise is observed as high energy at low frequency regions, we can observe high spectral energy density in the lower frequency regions (approximately below 40 Hz) for the genuine speech (Panel I) which is interestingly absent for spoofed speech (Panel II) for both STFT- and CQT-based

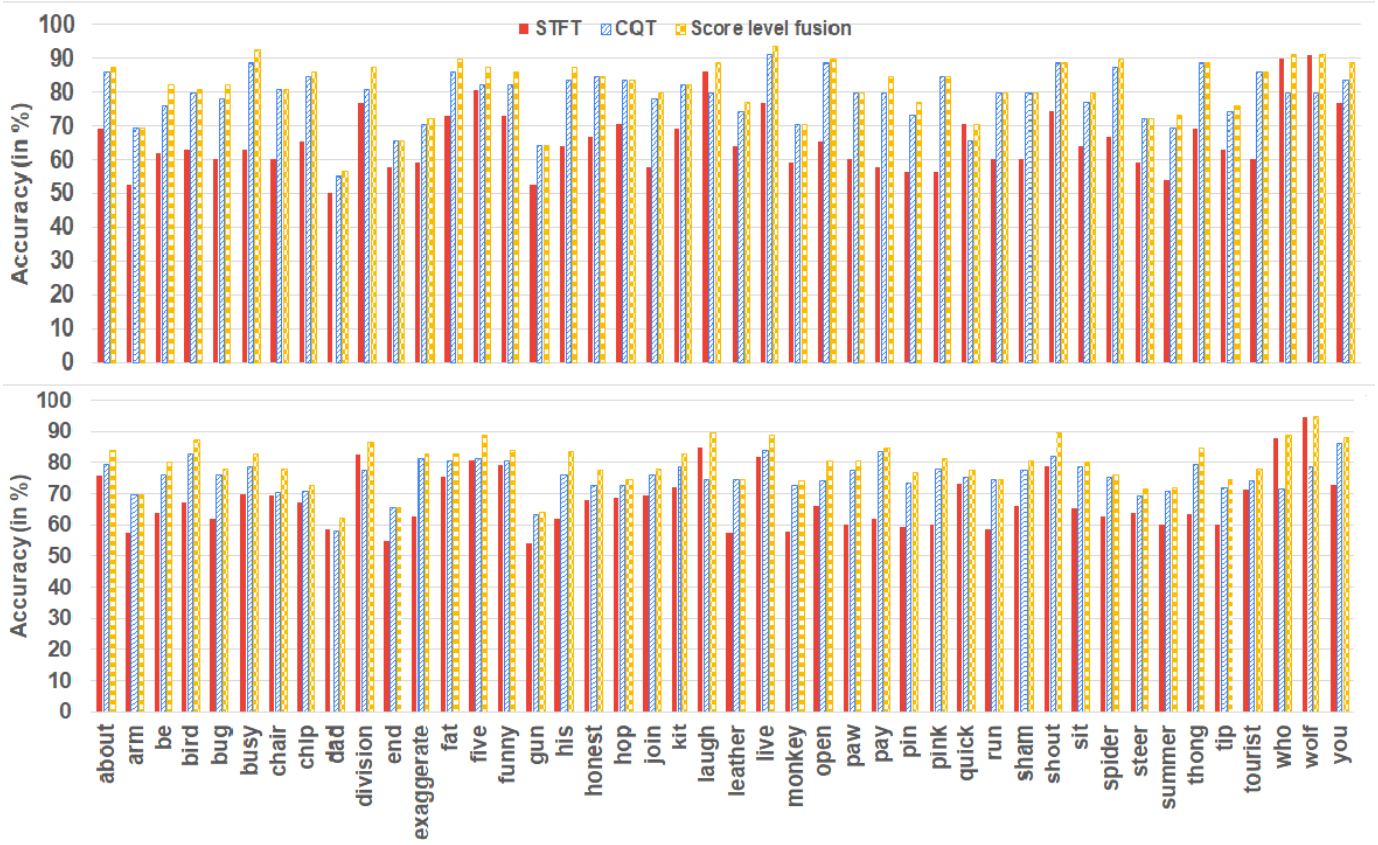

Fig. 3. Comparison of word accuracy on development and evaluation set for baseline *vs.* proposed algorithm. The word labels in Fig. 3(a) are similar to that of Fig. 3(b).

spectrograms. Furthermore, it can be observed that the spectral energy density obtained for pop noise using STFT spectrogram is relatively poor than CQT spectrogram, indicating that CQT brings out effect of pop noise more predominantly than the STFT. It might be due to high resolution characteristics of the CQT at low frequency regions. Hence, CQT is much better choice to obtain the higher classification accuracy.

B. Results

We have performed *10*-fold cross-validation and obtained overall accuracy of 62.29 % for baseline (STFT-based) and 66.49 % for proposed (CQT-based) algorithm. It can be observed that, the relative improvement of 6.4% in accuracy is obtained by the proposed algorithm over the baseline. We have also performed experiments for score-level fusion of the baseline and proposed algorithm. Fig. 3 (a) represents the comparison of word accuracy on development set for baseline, proposed algorithm, and their score-level fusion. Fig. 3 (b) represents the comparison of word accuracy on evaluation set for baseline, proposed algorithm, and their score-level fusion. Here, for baseline and proposed algorithm, we can observe that for words, such as 'busy', 'fat', 'funny', 'five', 'thong', 'shout' the accuracy is around 80 % as there is higher probability of presence of pop noise for these words. Moreover, proposed method performs much better as it gives higher accuracy for these words when compared to the baseline. For the other words, the accuracy is bit lower, still when compared the proposed methods, perform well than the baseline except for few words like 'laugh', 'who', and 'wolf'. Furthermore, the score-level fusion is performed on the likelihood scores obtained from baseline (STFT-based) and proposed (CQT-based) algorithm. We can observe that there is relative increase in % accuracy for both development and evaluation set for all the words. It suggest that the STFT and CQT captures the complementary information for liveness detection.

V. Summary and Conclusions

In this paper, we exploited the the pop noise as attribute of the genuine speech and it is effectively used for replay spoof attack detection. We proposed the novel approach of liveness detection using the features derived from CQT spectrogram on recently proposed POCO dataset. The results of proposed approach are compared against the baseline, where feature sets are derived from STFT. The spectrographic analysis for genuine (live) *vs.* spoof speech is performed which showed that the pop noise is emphasized in much better way for CQT-based spectrogram. The fact is validated by performing the experiments using *10*-fold cross-validation which shows that the proposed approach shows relatively better performance in detecting the liveness in speech. However, this approach

of replay spoof detection is useful only when the distance between microphone and speaker is less for genuine recording because pop noise can be captured from short distance only. Moreover, in the dataset it is assumed that the distance between attacker's recording device and genuine speaker is large and hence, the pop noise do not get captured by the recording device. In the future, it can be interesting to analysis the behaviour of pop noise when distance between attacker's recording device and genuine speaker is less i.e. the effect of distance between the speaker and recording device. Also replay of speech can be considered instead of using pop filtered speech to see the significance of pop noise from the view of practical deployment.

ACKNOWLEDGMENT

The authors would like to thank the authorities at DA-IICT, Gandhinagar for providing resources and kind support towards the completion of this research work.

REFERENCES

[1] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*. Springer, 2014, pp. 125–146.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

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

[5] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[6] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 195–202. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-28

[7] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," *TARGET*, vol. 4000, p. 5000, 2014.

[8] 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 *INTERSPEECH, Dresden,Germany*, 2015.

[9] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop,*, Les Sables d'Olonne, France, 26 - 29 June, 2018.

[10] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH 2019*, pp. 1008–1012. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2249 Last acessed date - 26-jan-2021

[11] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 681–685.

[12] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric." in *INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 671–675.

[13] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 82–86.

[14] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 1033–1037.

[15] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *INTERSPEECH*, Graz, Austria, 2019, pp. 1013–1017.

[16] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble Models for Spoofing Detection in Automatic Speaker Verification," in *INTERSPEECH*, Graz, Austria, 2019, pp. 1018–1022.

[17] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge." in *INTERSPEECH, Graz, Austria*, 2019, pp. 1038–1042.

[18] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing*. Springer, 2019, pp. 321–361.

[19] K. Akimoto, S. P. Liew, S. Mishima, R. Mizushima, and K. A. Lee, "Poco: a voice spoofing and liveness detection corpus based on pop noise," *in INTERSPEECH, Shenghai, China*, pp. 1081–1085, 2020.

[20] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. 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.

[21] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verifcation," in *Odyssey 2018 The Speaker and Language Recognition Workshop. ISCA, Les Sables d'Olonne*, 2018.

[22] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. 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.

[23] S. Mochizuki, S. Shiota, and H. Kiya, "Voice livness detection based on pop-noise detector with phoneme information for speaker verification," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 4, pp. 3060–3060, 2016.

[24] J. 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.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.