Morlet Wavelet-Based Voice Liveness Detection using Convolutional Neural Network

Priyanka Gupta, Piyushkumar K. Chodingala, and Hemant A. Patil
Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India
Email: {priyanka_gupta, piyush_chodingala, hemant_patil}@daiict.ac.in

***Abstract*—Given the attacker's freedom of using any spoofing attack, there is a need to explore liveness detection approaches that can classify a live speech from all the various spoofed speeches. To that effect, we propose Morlet wavelet-based approach for Voice Liveness Detection (VLD). We use acoustic cues of pop noise to discriminate a live speech signal from a spoof speech. Pop noise is present in live speech signals at low frequencies, caused by human breath reaching at the closely-placed microphone. As compared to the STFT-based baseline with 62.08% as overall accuracy, we obtain significantly improved performance. We achieve an overall accuracy of 80.00% on the evaluation set with *45*-D handcrafted Morlet wavelet-based features, and an accuracy of 86.23% with Morlet scalogram is obtained on the evaluation set. Better results signify that for VLD, wavelet transform-based time-frequency (scalogram) representation is more efficient as compared to the conventional STFT-based spectrogram. Furthermore, we have analyzed the effect of various phoneme types on VLD performance for the proposed approach.**

***Index Terms*—Voice liveness detection, automatic speaker verification, pop noise, scalogram, CNN.**

I. Introduction

An Automatic Speaker Verification (ASV) or voice biometric system performs machine-based authentication of speakers using speech signals [1]. ASV is a voice biometric system which has applications, such as banking transactions using mobile phones. Personal information, and banking details, demand more robust security of ASV systems. However, ASV systems are also vulnerable to various spoofing attacks, such as impersonation, twins, Voice Conversion (VC), Speech Synthesis (SS), and replay. Given these ASV system vulnerabilities, the ASVspoof Challenge campaigns were held in 2015, 2017, 2019, and 2021 during INTERSPEECH conferences. These challenges aimed to detect and develop robust countermeasures for replay attacks [2]–[6]. In this context, a 'liveness' detection corpus called as the POp noise COrpus (POCO) has been released in 2020 to allow research on development of robust Voice Liveness Detection (VLD) systems [7], [8]. VLD improves the security of ASV systems by protection against various types of spoofing attacks, including even attacks using unknown voice conversion and speech synthesis methods [8], [9]. One of the cues of liveness in a speech signal is the presence of *pop noise* in a live (genuine) speech signal. Pop noise is a short-time distortion in a speech signal which is caused by a burst of air on the microphone originating from a live speaker's mouth [10]. Signals that are known to spoof ASV systems, such as synthetic speech and replayed speech, fail to reproduce the pop noise as strongly as a live speech signal [7], [11], of course with the assumption that spoofed speech is not recorded with *wiretapping*. Pop noise is found in live speech as sudden bumps of strong energy within duration ranging between 20 ms and 100 ms [8]. This gives us a clue to define a suitable strategy for liveness detection. Low frequency regions ($\leq$ 40 Hz) are the regions where pop noise can be located [7], [8].

Considering the actual procedures for spoofing attacks, such as replay, SS, VC etc., spoofed speech has to be played via loudspeakers. Specifically for a replay attack, the recorded (spoofed) signal is captured in a covert manner from the genuine (live) speaker. Hence, the recording device is kept at a distance away from the speaker's microphone. The pop noise is poorly captured in the spoofed signal because of the distance of the recording device from the microphone and the assumption that no wiretapping is done. Moreover, playback devices and loudspeakers fail to reproduce the sudden distortions caused by pop noise [7], [12]–[14].

This paper Morlet wavelet-based features for pop noise detection. With respect to Heisenberg's uncertainty principle in signal processing framework [15], wavelet-based approach offers improved resolution in time and frequency as compared to the STFT-based method [16]. Furthermore, Morlet wavelets are known to capture perceptual cues effectively (both in visual and hearing domains). To that effect, the use of Morlet wavelet to capture discriminating cues based on pop noise for genuine *vs.* replay spoof classification is being proposed for the first time in this paper. Experiments are presented for two CWT-based features, namely, Handcrafted Morlet Wavelet and Low Frequency Morlet Scalogram-based Features on POp noise COrpus (POCO) for liveness detection.

II. Proposed approach using CWT

A. Continuous Wavelet Transform (CWT)

The effect of human breath on a microphone results in a sudden high energy (i.e., pop noise as an event in speech) in low frequency regions. To locate pop noise, time-frequency representations, such as spectrogram have been used in the past [8], [17]. However, to get better detection of pop noise, we have used CWT in this work. The key idea behind employing

ISBN: 978-1-6654-6798-8

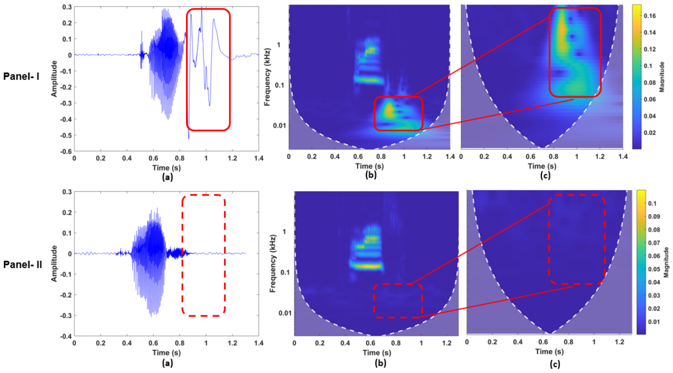

Fig. 1: Panel I represent the case of presence of pop noise (genuine speech) indicated by box. Panel II represents the case of reduced pop noise (spoofed speech) due to the use of pop filter, (a) time-domain signal for the word *'laugh'*, (b) corresponding scalogram, and (c) corresponding low-frequency ($0 - 40$ Hz) scalogram. Solid boxes in Panel I indicate the presence of pop noise, while corresponding dotted boxes in Panel II indicates that the pop noise has been eliminated due to pop filter.

wavelet for pop noise detection is to exploit the capability of a wavelet (which is a wave of short duration) to capture *transients* in speech, i.e., occurrence of pop noise. A mother wavelet $\psi(t) \in L^2(R)$ is a wave of short duration that has zero average. It is defined as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\Big(\frac{t-b}{a}\Big), \qquad a \in R^+, b \in R, \tag{1}$$

where b is called the translation (position), and a is called the dilation (scale) coefficient. There are various types of wavelets. The most famous wavelet is the Morlet wavelet, which is a modulated Gaussian, and it is defined as [18]:

$$\psi(t) = e^{j\omega_0 t}e^{-t^2/2}, \tag{2}$$

where ω_0 is taken as 5 Hz for a standard Morlet wavelet. The Morlet wavelet is obtained from a Gaussian window multiplied by a sinusoidal wave [19]. The CWT of signal $f(t)$ is

$$\begin{aligned} W_f(a,b) &=< f(t), \psi_{a,b}(t) >, \\ &= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\psi^*\Big(\frac{t-b}{a}\Big)dt, \end{aligned} \tag{3}$$

where $< \cdot, \cdot >$ indicates inner product operation to compute wavelet coefficients, and $*$ denotes complex conjugate. We have considered Morlet wavelet in this work because it is closely related to human perception (for both hearing and vision) [20]. Moreover, CWT is related to constant-Q filtering- a short-time analysis performed by the peripheral auditory system. In particular, as per the original investigation by Flanagan [21], the wavelet function, for the mechanical spectral analysis performed by the basilar membrane in the human ear is given by $\psi(t) = (t\omega)^2 e^{-t\omega/2}$ [21]. Furthermore, Morlet wavelet is the most widely used wavelet for CWT and, in fact, the first wavelet of its kind in formal historical developments of wavelets in the geophysics literature for detection of transients and improving joint time-frequency resolution of seismic signals [22].

B. *Proposed Approaches*

The feature extraction for Spoofed Speech Detection (SSD) task is based on the hypothesis that both genuine and spoof utterances possess differences w.r.t. presence and absence of pop noise energy levels respectively. Fig. 1 shows the scalograms of the word 'laugh'. A distinct signature of pop noise can be seen in Panel I. However, the pop noise signature is not so distinct for the case when a pop filter was used as shown in Panel II of Figure 1.

1) Handcrafted Morlet Wavelet-based Features: CWT coefficients are extracted from the speech data of POCO corpus by taking Morlet as the mother wavelet. CWT coefficients are found for frequencies $\leq$ 40 Hz, as shown in Algorithm 1. Furthermore, to keep the dimension (D) of feature vector as 45 and also to extract the prominent energy of pop noise, the energies are arranged in descending order, and the highest *45*-D values are taken for extracting the 45-dimensional feature vector.

Algorithm 1: Proposed Handcrafted Morlet Wavelet-based Feature Extraction for VLD.

```
Input: Speech signal $f(t)$
Output: Feature
w_name='amor'                        // Taking Morlet wavelet
[cwt_coeffs, F] ← cwt(f(t), w_name)
/* Finding CWT coefficients for low frequencies      */
Low_F ← find ($0 < F \leq 40$ Hz)
  Low_coeffs ← cwt_coeffs (Low_F)
Pop_energy = abs (Low_coeffs)$^2$
/* Converting pop energy to a 45-D feature vector    */
dim ← 45
M=mean (Pop_energy)
SD=standard_deviation (Pop_energy)
k ← length (Low_coeffs)
while $k > 0$ do
    $i = 1$
    $Norm\_Pop(i) = \frac{Pop\_energy(i) - M}{SD}$
    $k - -$ , $i + +$
[sorted, index] ← sort (Norm_Pop, descending)
Feature ← Pop_energy (index(1:dim))
```

2) Low Frequency Morlet Scalogram-based Features: Scalogram is a visual time-frequency representation of CWT coefficients. In particular, it can be interpreted as a time-frequency energy density, $|W_f(a,b)|^2$ [19]. The time-frequency resolution of the wavelet transform depends on the frequency of the signal. At high frequencies, the wavelet reaches a high time resolution but a low frequency resolution. At low frequencies, high frequency resolution and low time resolution can be obtained. Since pop noise most likely occurs at frequency regions $\leq$ 40 Hz, scalograms are very well suited to extract energies at low frequencies because of the higher frequency resolution of scalogram at lower frequencies.

For our experiments, the lowest frequency bin is set at 1.9826 Hz. The scale factor between 2 consecutive bins is

1.0718. Therefore, the k^{th} bin index corresponding to 40 Hz is calculated as:

$$40 = (1.0718)^k * 1.9826. \tag{4}$$

Therefore, frequency region approximately below 40 Hz is found to be corresponding to the nearest integer $k = 44$ frequency bins. Taking bin index below $k = 44$, we get frequencies exactly below 41.9025 Hz. This is the region where the pop noise is located. To that effect, scalogram images are extracted only corresponding to 44 wavelet coefficients. Each scalogram image is of the size 512×512. These scalogram-based features are then fed as an input to the CNN.

III. Experimental Setup

A. Dataset Used

The dataset used is the POp noise COrpus (POCO) with speech data sampled at 22.05 kHz [7]. The POCO dataset consists of 3 parts which are described briefly as below:

Genuine utterances with microphone-A (RC-A): In this set, only one microphone (Audio-Technica AT4040) is used. The distance between the speaker and the mic is kept fixed as 10 cm. The utterances in RC-A correspond to genuine utterances, as they have pop noise.

Genuine utterances with microphone array (RC-B): This set of genuine utterances is captured with a microphone array comprising 15 microphones (Audio-Technica AT9903 microphones).

Replay utterances with microphone-A (RP-A): Like the RC-A set, this set also contains utterances corresponding to one microphone and the distance between the speaker and the microphone is 10 cm. However, a TASCAM TM-AG1 pop filter is used between the speaker's mouth and the microphone. Given the use of pop-filter in this case, this set is emulated and considered to be spoofed and specifically designed for pop noise detection. Therefore, the utterances in RP-A correspond to spoof utterances, as they have pop noise.

TABLE I: Statistics of the POCO dataset used. After [7].

Subset	Number of Utterances	Number of Speakers	
		Male	Female
Training	6952	13	14
Development	3432	6	7
Evaluation	6600	13	13

Out of the above-mentioned subsets of the POCO dataset, we have used RC-A and RP-A as live and replay utterances, respectively. The speech samples of these 2 subsets were partitioned into training (40% of the dataset), development (20% of the dataset), and evaluation sets (40% of the dataset). The detailed specifications of the partitions taken are shown in Table I.

B. Classifier Used

A Convolution Neural Network (CNN) or ConvNet [26], [27] is a neural network model that consists of one or more convolutional layers followed by a classification layer. The two wavelet-based approaches described in Section II-B, which yield matrices of sizes 45×45 and $3 \times 512 \times 512$, respectively. For our experiments, the CNN architecture (shown in Figure 2) consists of 3 convolutional layers (Conv1, Conv2, Conv3) followed by 3 Fully-Connected (FC1, FC2, and FC3). The output of Conv3 is fed to the FC1 layer. The output of the final FC3 layer provides a probabilistic output for classification. Sigmoid activation function used at the output of FC3, while ReLU activation function is used for all the hidden layers. Binary cross-entropy is used as the loss function and stochastic gradient descent algorithm is used as the optimization algorithm. The sequence and the number of layers in the CNN are kept the same for *45*-D handcrafted feature as well as scalogram. However, for the case of scalogram images of size 512×512, the input is convolved with a kernel of size 7×7 for Conv1 and 3×3 for Conv2 and Conv3. For the case of handcrafted *45*-D wavelet-based features, the input is convolved with a kernel of size 3×3 during the forward pass, with a stride of 1, and zero-padding of 1. A max-pooling layer with a kernel size of 3×3, and stride of 1 is used.

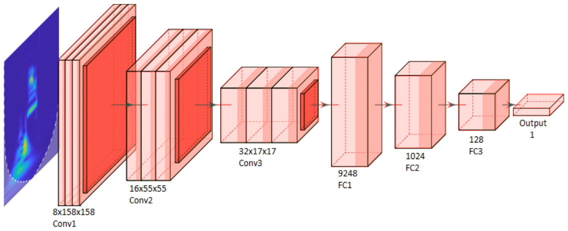

Fig. 2: The CNN architecture used for classification of the proposed Morlet wavelet scalogram-based features.

TABLE II: 44 words of the POCO dataset divided into phoneme categories. After [7].

Phoneme	Associated words in the dataset
Plosive	paw, tip, pink, open, pay, pin, sit, spider, be, kit, bird, end, dad, steer, quick, about, tourist, bug, honest
Fricative	wolf, laugh, five, funny, fat, live, shout, chair, sham, leather, thong, busy
Whisper	who, hop, you, his
Nasal	arm, monkey, summer
Liquids	run, gun
Affricate	chip, join, exaggerate, division

C. Phoneme-wise Categorization

There are 44 words in POCO dataset and their corresponding International Phonetic Alphabet (IPA) have been mentioned in [7]. Given that, a word can have multiple phonemes in it, only the most *prominent* phoneme in the word is taken into consideration. The 44 words of the POCO dataset are put into various phoneme classes as shown in Table II.

D. Baseline Approaches

1) Low Frequency Spectrogram-Based Features: Low frequency spectrogram-based features for VLD were extracted

TABLE III: Average Accuracy (in %) of different Phoneme types.

Phoneme Type	**(A) Spectrogram (SVM) [7]**	**(B) CQT (SVM) [23]**	**(C) Spectrogram (CNN) [24]**	**(D) Mel-spectrogram (CNN)**	**(E) Handcrafted Bump Wavelet-based (CNN) [25]**	**(F) Handcrafted Morlet Wavelet-based (CNN) (Proposed)**	**(G) Handcrafted Morlet Scalogram (CNN) (Proposed)**
Freq. Range	**0-40 Hz**	**0-11025 Hz**	**0-11025 Hz**	**0-40 Hz**	**0-40 Hz**	**0-40 Hz**	**0-40 Hz**
Plosive	60.46	63.61	71.72	74.13	81.58	79.35	**89.07**
Fricatives	67.66	73.78	75.55	77.45	80.77	79.27	**87.61**
Whisper	68.44	73.29	76.83	74.99	81.09	79.48	**86.21**
Nasal	54.26	57.78	59.33	70.51	76.50	71.36	**80.77**
Liquids	69.78	57.16	56	69.23	69.87	65.38	**79.49**
Affricates	58.26	68.92	71.83	72.51	78.53	74.35	**85.26**

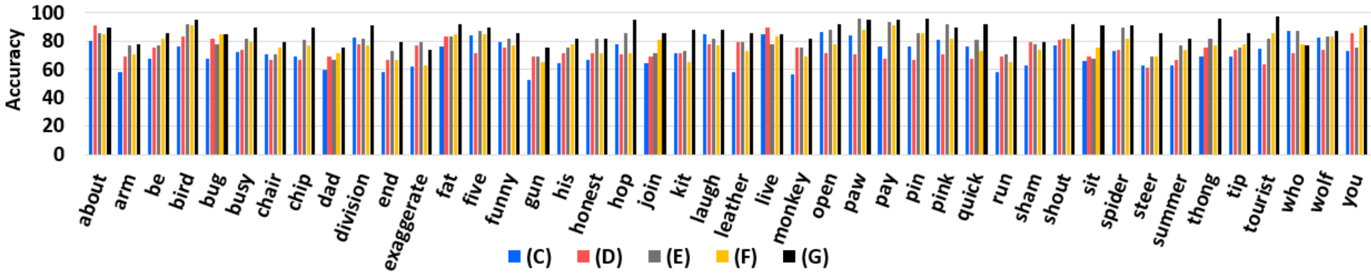

Fig. 3: Word-wise accuracies (in %) with CNN classifier for (C): Full-frequency spectrogram, (D): Low-frequency Mel-spectrogram, (E): Handcrafted Bump wavelet-based features, (F): Handcrafted Morlet wavelet-based features, and (G): Handcrafted Morlet scalogram.

from STFT in [8]. The same algorithm was used on POCO dataset in [7]. In this work, energies only in the low frequency (in particular, < 40 Hz) regions were extracted by selecting frequency bins corresponding to 0 to 40 Hz. Next, the average S_{eng} of the spectral energy densities of the STFT-based spectrogram was calculated by averaging across the bins for every k^{th} frame. For the framewise spectral energies obtained in S_{eng}, mean and standard deviation were calculated to obtain normalized values. The frames with the 10 highest energies were selected to get meaningful spectrogram-based features for pop noise detection. The classifier used was SVM.

2) CQT-Based Features: An improvement to the baseline was introduced in [23], using CQT-based features. As compared to the STFT that has constant frequency interval, CQT has geometrically distributed frequency bins due to constant-Q ratio of center frequency to resolution. The number of bins per octave is taken to be 96 and the number of samples taken in the first octave is 2. Furthermore, f_{min} is set to 0.48 Hz and f_{max} is set to 11050 Hz. For classification, the study reported in [23] used Support Vector Machines-based (SVM) classifier.

3) Mel-Spectrogram-Based Features: Apart from our proposed CWT-based approach in this work, we also include the use of Mel spectrogram (to our knowledge, this is not utilized for VLD task in the literature) for the purpose of comparing our results. We estimated pop noise energies using the STFT-based approach on Mel Spectrogram only on frequencies < 40 Hz. Therefore, we estimated the Mel-spectrogram with 16 number of bands and 5400 as the FFT length for better frequency resolution. Classification was done using a CNN-based classifier described in sub-Section III-B.

IV. Experimental Results

A. Proposed Handcrafted Morlet-Based Features

For the case of *45*-D wavelet-based features (shown as system (F)), we achieved an overall accuracy of 80%. Fig. 3 shows word-wise accuracy over 44 words in the dataset. We observed that the word 'pay' has the highest accuracy of 91.02%, because the word 'pay' has a strong plosive sound of /p/. Furthermore, we achieved an average accuracy of 79.35% and 79.27% on words with prominent performance on plosives and fricatives, respectively, as shown in Table III.

B. Proposed Morlet Scalogram-Based Features

The Morlet scalogram features (shown as system (G)) performed significantly well as compared to the traditional STFT-based baseline system. We observed overall accuracy of 86.23% on Morlet scalogram-based features. We observed that the word 'tourist' has the highest accuracy of 97.43%, because the word 'tourist' has 2 strong plosive sounds of /t/. Given the effect of pop-noise depends on the uttered word, we achieved an average accuracy of 89.07% and 87.61% on words with prominent plosives, and prominent fricatives respectively.

C. Discussion

It can be observed in Table III that our proposed Morlet scalogram-based approach outperforms every other methods for *all* the phoneme types. Furthermore, we also observe that all the methods perform relatively better for plosives and fricative sounds. Fricative sounds (such as, /f/ sound in the word 'lau**gh**') are produced due to turbulent airflow, which results in bursts of energy at low frequencies for a short-time period, characterizing the presence of pop noise. Furthermore, plosive sounds (such as, /p / sound in '**p**ay') are caused by a sudden release of a burst of air from the lips, resulting in pop noise [28]. On the contrary, energy distribution in nasal sounds is due to partial air released from the nostrils and the mouth [28]. Since the released air is coming from two sources, it barely results in energy at low frequency regions. To that effect, the accuracy score of all the algorithms are relatively lower for the nasal sounds.

V. Summary and Conclusions

In this work, we used CWT to effectively improved resolution in time and frequency for VLD based on pop noise. VLD enables to discriminate a *live* voice from the other *non-live* voice signals, such as replayed, voice converted, and synthetically generated signals. To that effect, two handcrafted features were proposed in this study: Morlet wavelet-based features, and Morlet scalogram-based features. A significant improvement in accuracy is observed with both the features as compared to the existing systems. Further analysis shows the effect of phoneme type on the accuracy. However, the proposed approach comes with a trade-off between high performance and computational complexity. Further similar wavelet-based methodologies can be tested for various configurations of spoof signals, as future work. Furthermore, the combined effect of microphone variability on ASV and pop noise-based VLD task can also be investigated.

VI. Acknowledgements

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] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] W. Zhizheng and et. al., "ASVSpoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in 16th *Annual Conf. of the Int. Speech Communication Association*, Dresden, Germany, 6-10 September, 2015, pp. 2037–2041.

[3] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVS 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 7–11.

[4] M. Witkowski and et. al., "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 20-24 August, 2017, pp. 27–31.

[5] J.-w. Jung, H.-j. Shim, H.-S. Heo, and H.-J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVSpoof 2019 challenge," *arXiv preprint arXiv:1904.10134*, 2019, {Last Accessed: 2020-01-31}.

[6] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVSpoof 2019," in *IEEE Autom. Speech Recognit. Understanding (ASRU) Workshop, Singapore*, 2019, pp. 1018–1025.

[7] 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," *INTERSPEECH*, pp. 1081–1085, Shanghai, China, 25-29 Oct., 2020.

[8] 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, 6-10 September, 2015, pp. 2047–2051.

[9] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu, "Secure your voice: An oral airflow-based continuous liveness detection for voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, NY, USA*, vol. 3, no. 4, pp. 1–28, 2019.

[10] 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.

[11] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, 2017.

[12] Y. Hsu, "Spectrum analysis of base-line-popping noise in MR heads," *IEEE Transactions on Magnetics*, vol. 31, no. 6, pp. 2636–2638, 1995.

[13] Shiota, Sayaka and Villavicencio, Fernando and Yamagishi, Junichi and Ono, Nobutaka and Echizen, Isao and Matsui, Tomoko, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector." in *Speaker Odyssey, Bilbao, Spain*, vol. 2016, 2016, pp. 259–263.

[14] 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, pp. 233–239.

[15] D. 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.

[16] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[17] 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 Conference on Computer Communications*, 29 April - 2 May 2019, Paris, France, pp. 2062–2070.

[18] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

[19] S. Mallat, *A Wavelet Tour of Signal Processing, 2nd Ed.* Elsevier, 1999.

[20] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[21] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer Science & Business Media, 2013, vol. 3.

[22] I. Daubechies, "Where do wavelets come from? A personal point of view," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 510–513, 1996.

[23] K. Khoria, Ankur T. Patil, and Hemant A. Patil, "Significance of Constant-Q transform for voice liveness detection," in *EUSIPCO*, Dublin, Ireland, 23-27 August 2021.

[24] S. Gupta, K. Khoria, A. T. Patil, and H. A. Patil, "Deep convolutional neural network for voice liveness detection," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), TOKYO, JAPAN.* IEEE, 2021, pp. 775–779.

[25] Priyanka Gupta, Siddhant Gupta, and Hemant A. Patil, "Voice Liveness Detection using Bump Wavelet with CNN," in *International Conference on Pattern Recognition and Machine Intelligence LNCS.* Springer, 15-18 December, 2021.

[26] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1, no. 2.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. First Edition, Pearson Education India, 2006.