Modified Group Delay Cepstral Coefficients for Voice Liveness Detection

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

***Abstract*—Liveness detection has advanced for many biometrics, such as face, iris, hand geometry, etc. However, less emphasis is given to liveness detection for voice biometrics, i.e., Voice Liveness Detection (VLD). Pop noise is produced due to the spontaneous breathing while uttering a certain class of phonemes (such as fricative, affricates, plossive, nasal, etc.) which has low frequency characteristics. In this paper, phase-based information for analysis of pop noise is used as a feature for VLD task. We have use modified group delay function based cepstral coefficients (MGDCC) as a feature for VLD task. Experiments are performed using two different types of classifiers, i.e., Gaussian Mixture Model (GMM) and Convolutional Neaural Network (CNN). Our results indicate an overall improvement in accuracy by 17.16% and 1.38% for MGDCC-CNN and MGDCC-GMM systems respectively.**

***Index Terms*—Voice liveness detection, Modified Group Delay Function, Pop noise, GMM, CNN, POCO dataset.**

I. Introduction

Due to recent technological developments, voice biometrics have been extensively used for banking transactions, border control, homeland security, smartphones, etc. Hence, the need for robust Automatic Speaker Verification (ASV) system has also increased [1], [2]. The ASV system is used to prevent the voice biometrics from harmful intent of the attacker. However, with the recent progress in various speech technologies, attempts to spoof an ASV system has been made using spoofing attacks, such as voice conversion, speech synthesis, replay, and impersonation attacks [3]–[5]. The severity of these attacks has made the spoofing detection as one of the important research issue in the field of ASV. In the past five years, several challenges have been organized, such as ASVSpoof challenges during INTERSPEECH conferences with the aim to improve performance of anti-spoofing for attack-resistant design of ASV systems. [6], [7].

In the light of these challenges, many countermeasures to detect spoofing attacks have been developed and evaluated on the standard datasets provided by the ASVSpoof challenge organizers. These challenge campaigns focussed on performance of the countermeasure systems for the anti-spoofing. In this paper, the focus is centered on the Voice Liveness Detection (VLD) to avoid the spoofing attacks on ASV system. Recently, POCO (POp noise COrpus) dataset [8] is build to develop various countermeasures against spoofing attacks to detect the human liveness evidence in the speech signal via pop noise detection. Identification of the genuine speaker characteristics (i.e., VLD task) through pop noise detection can be potentially effective when the distance between the testing microphone and the speaker is very less, which consequently leads to detect the spoofing attacks.

Human liveness detection for detecting spoof attacks was proposed for the first time in [9], where two approaches for VLD are proposed: (a) low-frequency based single channel detection, and (b) subtraction-based pop noise detection with two channels. In the former approach, entire 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 [10], phoneme-based pop noise detection is performed for VLD along with ASV system, where pop noise duration is detected in the utterance and estimated phonemes in this duration are analyzed for VLD. The similar approach of phoneme-based pop noise detection was utilized in [11] with extended study on Gammatone Frequency Cepstral Coefficients (GFCC) feature set for pop noise detection.

To the best of the authors' knowledge, this is the first study of its kind to detect pop noise by exploring the phase-based features which exploits Fourier transform phase spectrum information of the signal rather than the conventional magnitude spectrum-based features. We have applied Modified Group Delay (MGD) funtion as feature extraction method which provides discriminative information in spectral regions by showing better spectral resolution in comparison to the magnitude spectrum [12], [13]. Furthermore, the feature obtained from MGD function is used with GMM and CNN classifier to detect the presence of pop noise in a signal, which has shown satisfactory results.

II. Proposed Algorithm

A. Pop Noise

During natural speech production mechanism, speech wave is a result of airflow travelling from the lungs to the vocal folds which excites the vocal tract system (that acts as cascade of several 2^{nd} order resonators representing organ pipes) causing the bursts of air coming out of the mouth [14]. If this sound is captured at a small distance from the microphone and a speaker, the microphone along with the speech signal also tends to capture the friction between the lips as *bursts* which is termed as *pop noise* [9]. The distance between the speaker

and microphone and the intensity of the pop noise detected via microphone have Inverse relationship with each other. The intensity of the recorded pop noise cannot be high if the distance between the microphone and the speaker is large. This phenomena can be used to prevent the information in speech signal from the attacker who is trying to record the voice for fraudulent attack. The attacker may not be able to place the recording device near the speaker leading to the absence of pop noise in the recorded voice. Hence, detection of *pop noise* can provide genuine acoustic cues for VLD which will further be able to detect between the live (genuine) speech and replayed speech.

B. Baseline Algorithm

In our work, detection of pop noise is considered as two-class classification task, where utterances with pop noise are labelled as genuine while utterances with absence of pop noise is labelled as imposter (spoof). The baseline is implemented with the methodology given in [8]. Spectrograms are used as input features.S_{eng} is obtained by considering spectral energy densities corresponding to [0, f_{max}]. Since the pop noise is observed in the lower frequency region of the spectrogram features, f_{max} is considered as $40Hz$. After that f_{avg} is estimated as the average of the spectral energy densities for each frame. Then, the mean and standard deviation is taken for f_{avg} across all the frames. Then, mean and standard deviation is estimated for averaged spectral energies f_{avg} to normalize it. Then, 10 frames with largest spectral energies were chosen. This is done by taking 10 frames from normalized f_{avg} having largest value and then taking frames corresponding to that indices from S_{eng}. This feature set with appropriate labels, is given as input to Support Vector Machine (SVM) for proposed/VLD task. The more details of this baseline algorithm is given in [9].

C. MGD feature

It is common practice to extract information from the magnitude spectrum of the signal since it involves less computational complexity as compared to phase spectrum which requires computationally intensive task of phase unwrapping to invert artifacts of arctangent function. However, the speech signal information resides in both the magnitude and phase component which has motivated to explore the phase characteristics of the pop noise signal including very recent application in spoofed speech detection [15]. To this initial step, we have considered Group Delay (GD) function which observes the rapid variations in the unwrapped phase function and has a property to better resolve the resonant peaks of a signal. GD function is a potential alternative to the magnitude spectrum and has been employed extensively in a feature extraction process for detecting other spoofing attacks along with conventional classifiers such as Gaussian Mixture Models and complex Deep Neural Network based classifiers [16]–[20].

The true meaning of the GD function lies in the time shifts introduced by the systems having linear phase characteristics. This concept can be simply extended to the non-linear phase characteristics by linearly approximating the narrowband input of the non-linear phase system. The approximate effect of the system on the input involves magnitude shaping and multiplication by complex constant factor and localized linear phase term corresponding to time delay (in seconds). This overall time delay introduced by the system in the input is referred as Group Delay [21] i.e., this group delay is being experienced by a group of frequencies (narrowband corresponding to localized linear phase slope) in the input.

Fig.1 is shows speech segments, magnitude spectrum, group delay function, and modified group delay function for both the classes of speech. We can see that group delay function is showing very poor speech structure. This is primarily because the speech segments considered are non-minimum phase signal in which presence of zeros near the unit circle in the z-plane causes sharp spikes, interfering with the speech structure while MGD improves the structure (to be discussed shortly). In additional, better speech structure formation for pop noise through MGD can be seen than without pop noise speech segment.

1) Feature Extraction: The speech signal $x(n)$ is analyzed with the help of Short-Time Fourier Transform (STFT) to extract the MGD feature. The magnitude and phase representation of $x(n)$ is given by Eq.(1):

$$X(\omega) = |X(\omega)|\, e^{j\phi(\omega)}, \tag{1}$$

where $|X(\omega)|$ and $\phi(\omega)$ are the magnitude and phase spectrum at frequency ω. To extract the information from the STFT phase function of the speech signal, negative derivative of the phase function is processed which is the Group Delay (GD) function, i.e. ,

$$\tau(\omega) = -\frac{d}{d\omega}\phi(\omega) = -imag[\frac{d}{d\omega}log(X(\omega))], \tag{2}$$

In Eq.(2), GD function requires the phase function $\phi(\omega)$ to be unwrapped which satisfies that all the multiples of 2π have been included in $\phi(\omega)$ to contribute towards the true time delay which is a complex task [22]. Therefore, the GD function is computed by invoking the Fourier Transform(FT) property of instantiation frequency domain as shown in Eq.(3):

$$\tau(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{|X(\omega)|^2}, \tag{3}$$

where $X(\omega)$ and $Y(\omega)$ are the STFT of the $x(n)$ and $nx(n)$, r and i are the real and imaginary parts respectively. The GD function introduces effects, namely, spikes and pitch periodicity which occurs when zeros of $X(\omega)$ lie close to the unit circle in z-plane which causes denominator term to become very small. The MGD was introduced to avoid these effects by cepstral smoothing the denominator $|X(\omega)|$ term with the help of parameters ρ and γ [12], [13], [23], [24]. The MGD function is represented in Eq.(4) as follows:

$$\tau(\omega) = \frac{X_r(\omega)Y_r(\omega) + X_i(\omega)Y_i(\omega)}{|X_c(\omega)|^{2\rho}}, \tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\gamma}, \tag{4}$$

where $|X_c(\omega)|$ is the cepstrally smoothed version of $|X(\omega)|$ and ρ and γ are the fine tuning parameters. $\tau_m(\omega)$ is the final MGD function. The algorithm for the MGD function is given as follows [13] :

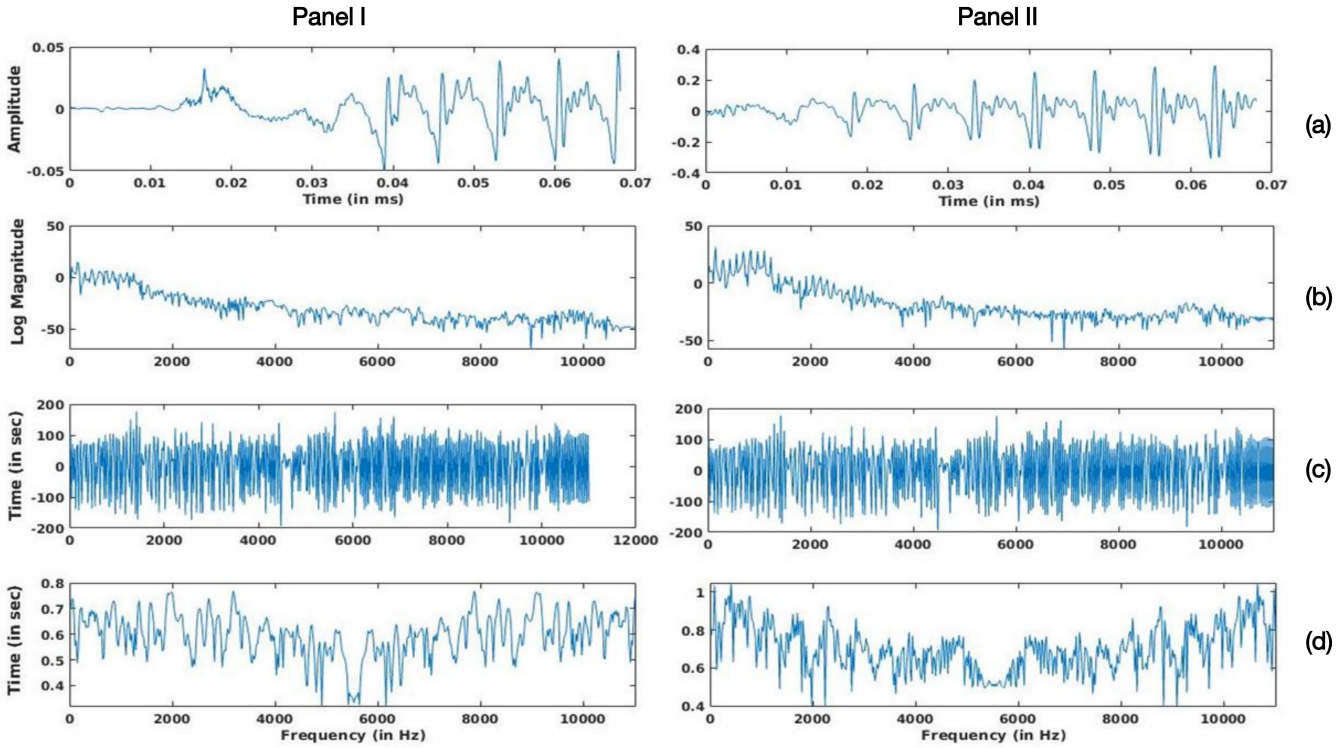

Fig. 1. Panel I (a), (b), (c), (d) shows the speech segment containing pop noise (genuine) and corresponding magnitude spectrum, group delay, and modified group delay function, respectively, and corresponding plots for without pop noise (spoofed) utterance is shown in Panel II.

- Compute the STFT of the signal $x(n)$ and $nx(n)$ i.e., $X(\omega)$ and $Y(\omega)$.
- Perform the cepstral smoothing on $|X(\omega)|$ in order to obtain $|X_c(\omega)|$.
- Compute the MGD function as given in Eq.(4).
- Compute the MGD function for different values of the parameter ρ and γ to get the better results.
- Applying Discrete cosine transform (DCT) to obtain cepstral coefficient.

III. Experimental Setup

A. Dataset Details

In practical scenarios, if an attacker tries to attempt a spoofing attack, he/she must somehow obtain the voice samples of the target (genuine) speaker. The simplest way to do this is by recording (eavesdropping) the voice of target speaker and then using it to mount a replay attack onto the ASV system. Since these recordings will be done from long distances, pop noise will not be recorded by the attacker's microphone and this absence of pop noise in the replayed sample will be able to *flag* the spoofed speech from the genuine speech.

In this work, we have used recently released *POCO* dataset [8]. There are total of 66 speakers out of which 34 are male and 32 are female. The words were selected from the English language such that all the 44 phonemes are covered in the recording. The dataset is sampled at 22050 Hz sampling frequency with a bit-depth of 24-bits. 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. In addition, the experiments in [8] are performed using RC-A and RP-A subsets. The details of RC-A and RC-B are as follows:

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

2) Eavesdropping (RP-A): Eavesdropping is done to imitate a scenario where replay attack is done by an attacker from a long distance, i.e., without pop noise. This condition is simulated by using Audio-Technica AT4040 microphone with a suitable pop filter inserted between speaker and microphone. The distance between speaker and microphone was fixed as 10 cm. The dataset is partitioned into training and evaluation subsets as 80% and 20% utterances, respectively. Each of these subsets consists 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 and evaluation subset is shown in Table I.

B. Feature Set

Modified Group Delay Cepstral Coefficient (MGDCC) is computed from the Modified Group Delay function of the speech signal converting them into cepstral coefficients. De-

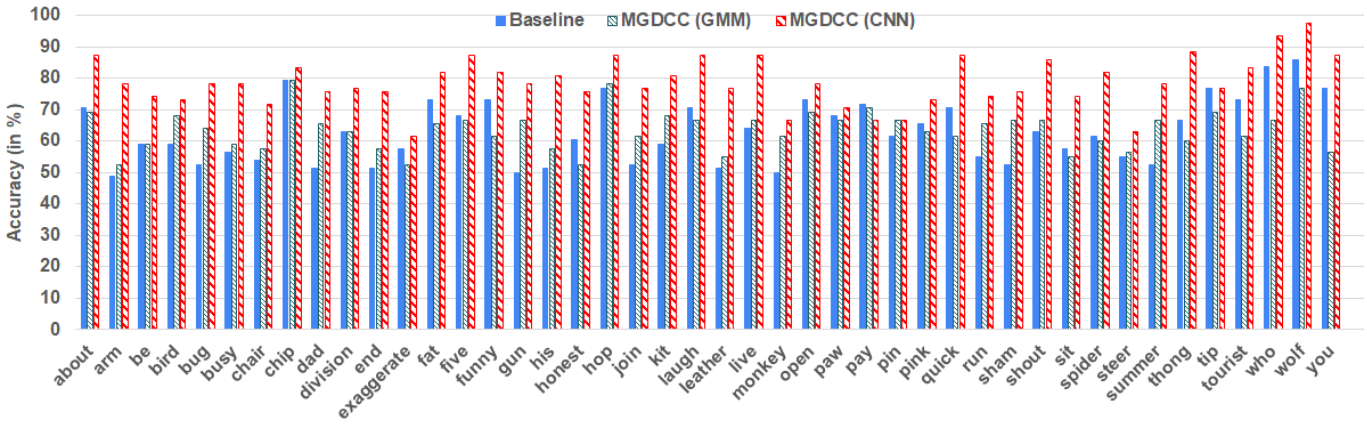

Fig. 2. Comparison of accuracy (in %) for baseline, MGDCC (GMM), and MGDCC (CNN)

TABLE I
STATISTICS OF THE POCO DATASET FOR OUR EXPERIMENTS

Subset	# Utterances	# Speaker	# Male	# Female
Training	13552	53	26	27
Evaluation	3432	13	6	7

tails about the computation and advantages of the MGDCC feature has been discussed in the Section II-C. The range of the fine tuning parameters, ρ and γ varies between 0 to 1 which can be fixed depending on the experimental analysis of the problem at hand. For the classification task, MGDCC feature has been extracted with the help of Hamming window of 25ms duration and 10 ms shift along with *16*-Dimensional cepstral features for various combinations of ρ and γ. Experimental results for the implementation of different values of tuning parameters are discussed in Section IV.

C. Classifier

We have performed experiments using Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN) as classifiers. GMM is used as a two-class classifier, where the two classes correspond to the speech samples containing pop noise (genuine) without pop noise (spoofed). The individual GMM is trained on genuine and spoofed speech using the extracted MGDCC feature sets. Total 128 GMM mixture models are used. The CNN network consists of 3 convolution blocks, and 3 Fully-Connected (FC) layers. Each convolution block consists of a $2-D$ convolution layer accompanied by a max-pooling layer to remove the inconsistencies in the feature map. Kernel size of 3×3 is taken for both convolution and max-pooling operations. In addition, convolution operation is performed using zero padding with a stride of 2. The final convolution block is followed by 3 fully-connected linear layers with distinct hidden units. Sigmoid is used as an activation function at the output of final layer to make the final decision of whether the utterance contains pop noise or not. In hidden layers, Rectified Linear Unit (ReLU) function is used as the activation function. 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 experiments are performed using speaker-independent and customized disjoint partition of the dataset as shown in Table I.

IV. EXPERIMENTAL RESULTS

In Fig. 2 wordwise accuracy is shown for baseline, MGDCC-GMM , and MGDCC-CNN, where GMM and CNN is used as a classifer. It can be observed that for words such as sham, shout, summer, bird and, gun MGDCC-GMM is performing slightly better than the baseline algorithm while for other words both are almost comparable. For few words, such as quick, who, wolf and, you the performance of MGDCC-GMM is slightly poorer than the baseline. The probability of presence of the pop noise in differents words vary which will also affect the system performance. For plosive sound presence of pop noise is expected to be more as it produces due to friction from the lips whereas words containing nasal sound will have less pop noise intensity. However, when CNN is used as a classifier with MGDCC feature set (MGDCC-CNN), there is significant improvement in the accuracy when compared to the baseline. The average accuracy for MGDCC-CNN system is obtained as 79.45 % which is 63.67 % for MGDCC-GMM system and 62.29 % for the baseline system.

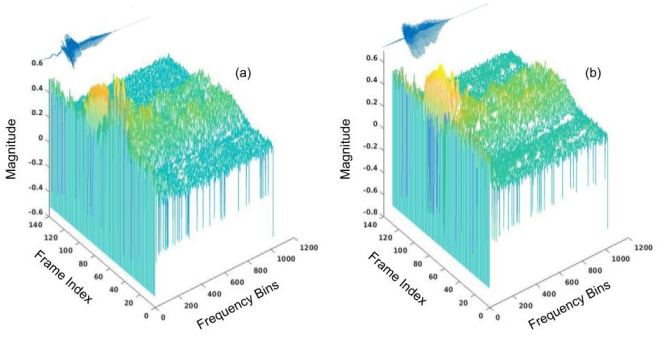

Fig. 3. Waterfall plot showing the differences between speech signals (a) containing pop noise, and (b) without pop noise.

Hence, their is improvement of around 17.16 % for MDGCC-CNN system and around 1.38 % for MGDCC-GMM system when compared to the baseline system. It can be also observed that when CNN is used as classifier along with MGDCC, the performance is improved when compared to MGDCC-GMM system. In additional, for MGDCC-CNN system for word which have high probability of pop noise, such as thong, shout, who, five, and wolf the average accuracy is around 85 % which is around 70 % for the baseline system. The results suggests it is necessary to consider the phoneme information of the utterances to reject spoofing attacks more robustly, which is introduced in [25].

In Fig. 3, waterfall plot is shown for genuine (with pop noise) and spoofed (without pop noise, i.e., with pop filter) utterances. Differences between the genuine signal and spoofed signal can be observed clearly, which is learned by the classifier for the VLD task. Also, for different combinations of tuning parameters, some of the best results observed is shown in Table II with GMM as a classifier. Particularly for this experiment for the better analysis, dataset is divided into 40% for the training dataset, 20% for the development, and 40% for the evaluation dataset making sure that the speakers are exclusive in each dataset.

TABLE II
RESULTS FOR MGDCC-GMM FOR SOME COMBINATIONS OF γ AND ρ.

γ	ρ	Accuracy (%) on Development set	Accuracy (%) on Evaluation set
0.1	0.4	77.45	70.05
0.2	0.2	79.90	71.20
0.2	0.8	75.55	69.15
0.4	0.8	69.81	65.53

V. SUMMARY AND CONCLUSIONS

This study presents one of the first study study to investigate the phase-based features to detect voice liveness. MGDCC feature alone has shown promising results using both the GMM and CNN classifiers. This has shown path to investigate other phase based features which can be combined with magnitude based features to extract signal information to a greater extent. Furthermore, fusion of the features at score-level can also be performed to get better insight on the pop noise signal characteristics. The future work will focus towards utilizing the phase-based features with deep learning architectures to detect pop noise.

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] 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

[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] 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.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXivreprint arXiv:1904.05441*, pp. 1008–1012, 2019.

[7] 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.

[8] 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, Shanghai, China*, pp. 1081–1085, 2020.

[9] 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.

[10] 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.

[11] 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, Paris, France*, 2019, pp. 2062–2070.

[12] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202.

[13] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1, Hong Kong,China, 2003, pp. I–68.

[14] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. 2^{nd} Edition, Pearson Education India, 2006.

[15] J. Yang, H. Wang, R. K. Das, and Y. Qian, "Modified magnitude-phase spectrum information for spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021.

[16] 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.

[17] K. Srinivas, R. K. Das, and H. A. Patil, "Combining phase-based features for replay spoof detection system," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Taiwan,China, 2018, pp. 151–155.

[18] K. Srinivas and H. A. Patil, "Relative phase shift features for replay spoof detection system," in *Proceddings of the Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, New Delhi,India, 2018, pp. 1–5.

[19] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Thirteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2012.

[20] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proceedings of The 2012 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference,(APSIPA ASC)*, 2012, pp. 1–5.

[21] A. V. Oppenheim, A. S. Willsky, and S. Hamid, "Signals and systems, processing series, 2^{nd} edition," 1997.

[22] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 170–177, 1977.

[23] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

[24] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on signal processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[25] 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.