Pop Noise Detection Using Group Delay Cepstral Coefficients

Arth Shah, Prathav Kevadiya, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

E-mail:{202101154, 202003020, hemant_patil } @daiict.ac.in

Abstract

This proposed study leverages the distinctive pop noise, which is involuntarily generated when pronouncing specific phonemes, such as plosive, fricative, nasal, and affricate sounds. We employ Group Delay Cepstral Coefficients (GDCC) as the key features for the Voice Liveness Detection (VLD) task. To measure the effectiveness of GDCC, we compared the performance using different spectral features, namely, Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Gammatone Frequency Cepstral Coefficients (GFCC), employing them with three types of classifiers, namely, Convolutional Neural Networks (CNN), Convolutional Recurrent Neural Networks (CRNN), and Localized Convolutional Neural Network (LCNN). We achieved an accuracy of 79.55% for the GDCC feature vector as its highest, which is comparatively much higher than other feature vectors, thereby proposing a new optimal approach for the VLD task.

1. Introduction

Robust security has become more critical than ever in today's digital world, where remote communication and online transaction systems dominate. Voice biometrics has gained popularity as it's a practical, effective, and natural method for verification of a person's identity. In the meantime, advancements in speech synthesis and transformation technologies have made it possible to produce realistic-sounding artificial (or spoofed) speech with the audio of the target speaker from a supplied text or an inputted audio waveform pronounced by an imposter speaker. Nonetheless, analogous methods might be employed to create counterfeit identities or profiles and initiate spoofing assaults on Automatic Speaker Verification (ASV) systems [1], [2], thereby presenting a substantial risk to the security of these systems. To that effect, researchers of this study focus on developing advanced Voice Liveness Detection (VLD) systems to distinguish between genuine *vs.* spoofed voices, which is more reliable.

Several studies have been explored to increase the reliability of VLD systems in recent years; some countermeasures used the ASVSpoof challenge datasets to develop and evaluate their model. Nevertheless, inspired by [3], we also employ pop noise in our work to protect ASV systems against spoofing assaults. In this context, a new dataset called POCO (POp noise COrpus) [4] has also been developed in the literature to make it easier to design a variety of countermeasures that seek to find pop noise—a marker of human liveness in voice signals. One typical type of speech distortion is called "*pop noise*," which occurs when air from a speaker's lips enters a microphone and is misrepresented by speakers. When assessing an authentication system, it seems sensible to use pop noise as a trustworthy signal of liveness, given its common occurrence. The ability to differentiate between genuine audio and replayed audio (through loudspeakers) may be effectively established by considering a pop noise check.

Phase-based features have become a potential alternative to magnitude spectrum-based features for several speech applications. Phase-based features have noise robustness, perceptual relevance, anti-spoofing ability, and discriminating power. These attributes help to differentiate between authentic human voices and spoof or prerecorded signals because they capture fine-grained *temporal* and *phase*-related properties of the speech signal. In this study, we exploit the Group Delay Function (GDF) for feature extraction to achieve better spectral resolution than the magnitude spectrum for the VLD task. Further, the features obtained from the GDF are combined with Convolutional Neural Networks (CNN), Convolutional Recurrent Neural Networks (CRNN), and Localized Convolutional Neural Network (LCNN) classifiers to detect pop noises, which are found to give promising results.

1.1. Pop Noise

While producing normal speech, a live speaker generates a speech distortion known as pop noise [5]. This phenomenon occurs when the Vocal Tract System (VTS) is stimulated by air from the lungs, causing the vocal folds to vibrate [6]. As the glottal airflow passes through a cascade of organ pipes, it reaches the mouth, where a specific phoneme being uttered determines the location and force of the burst of air, producing a sudden outburst of sound from the mouth. Pop noise may have a different intensity depending on how close the speaker is to the recorded microphone.

By detecting and analyzing pop noise, it becomes possible to protect speech signal information from potential attackers attempting fraud or unauthorized voice recordings. Detecting pop noise provides genuine acoustic cues for VLD, enabling distinguishing between live (natural) speech *vs.* replayed the speech. This distinction is crucial in preventing fraudulent activities, and enhancing the security of voice-based systems [7]. Therefore, investigating and identifying pop noise characteristics and developing techniques to detect and analyze them contribute to the development of effective VLD methods. By leveraging pop noise as an acoustic cue, VLD systems can distinguish between genuine live speech and spoofed or manipulated recordings, enhancing the reliability and security of voice-based authentication and verification systems. This study provides the following novelty:

- Mathematical as well as theoretical analysis of GDCC feature vector.
- Discussion of Results obtained.
- Experimental analysis related to the feature vector's evaluation factor and latency analysis.

- Comparison with different features on different classifiers, and existing approach.

The paper's remaining section is organized as follows: With a thorough theoretical exposition and a comparison of GDF with other features, Section 2 offers computational details on GDCC feature extraction. The dataset, performance measures, classifiers utilized in the experiments, and other spectral properties besides GDCC used for comparison are all described in depth in Section 3. Using the GDCC (proposed) feature set, the results are shown in Section 4 along with a comparison to previous research. Lastly, a summary of the work and recommendations for further research are provided in Section 5.

2. Proposed GDCC Features

The information offered by the magnitude spectrum is supplemented by the group delay spectrum, which provides insightful information about the *temporal* properties of speech signals [8]. The speech signal's envelope and fine structure are determined by the VTS and excitation source, respectively. The objective of extracting features from the magnitude spectrum involves identifying the spectral envelope of the audio signal. It is possible to gain more information about speech signals, and improve the effectiveness of specific speech analysis tasks by integrating *spectral* and *temporal* information.[8]

Because of its lower complexity compared to the phase spectrum, the magnitude spectrum facilitates the relatively straightforward extraction of features containing information about the signal, which requires addressing the intricate task of phase unwrapping in order to reverse the signal processing effects inherent in the arctangent function. [8]. Speech signals are a combination of magnitude and phase spectrum. This study explores phase-based characteristics of signals containing pop noise. In order to achieve this, we use the GDF, which, by performing a negative derivative operation on the unwrapped phase function, enables us to see quick fluctuations in it, i.e., the group delay function [9]. Better resolution of the resonant (peaks) formant is made possible by this GD function characteristic as opposed to the short-time speech spectrum.

Many researchers in the field of speech technology have employed magnitude-based acoustic features, resulting in neglecting of phase spectrum-based characteristics of speech signals. GDF is based on the shift initiated by a *group* of frequencies in a system containing linear phase characteristics [9]. A similar process can be applied to systems with nonlinear phase characteristics by linearly approximating their narrowband input. The input is affected by the system in three ways: it shapes its magnitude, multiplies it using a complex function, and adds a localized linear phase term that represents the delay. The study of GDF is concerned with the delay that a set of frequencies experiences when they are introduced into the system.[10].

For a discrete-time speech signal x(n), its Discrete-Time Fourier Transform (DTFT) represented as X($e^{j\omega_0}$), which is expressed via magnitude-phase representation as [5]:

$$X(e^{j\omega_0}) = \mid X(e^{j\omega_0}) \mid e^{j\phi(e^{j\omega_0})}, \tag{1}$$

where the phase spectrum at frequency ω_0 is represented by $\phi(e^{j\omega_0})$, and the magnitude spectrum is represented by $|X(e^{j\omega_0})|$. The negative derivative of the unwrapped phase, or GDF, has been studied in attempt to extract information from the phase spectrum. In particular,

$$\tau(e^{j\omega_0}) = -\frac{d}{d\omega_0}\phi(e^{j\omega_0}) = -j\,\mathrm{mag}\left[\frac{d}{d\omega_0}\log(X(e^{j\omega_0}))\right].$$

In order to avoid the computationally intensive task of phase unwrapping, an alternative way to compute the GD function is by exploiting the frequency-domain differentiation property of DTFT [11], [12]. In particular,

$$\tau(e^{j\omega_0}) = \frac{X_R(e^{j\omega_0})Y_R(e^{j\omega_0}) + X_I(e^{j\omega_0})Y_I(e^{j\omega_0})}{|X_c(e^{j\omega_0})|^2},$$

where $X(e^{j\omega_0}) \rightarrow$ DTFT of x(n), and $Y(e^{j\omega_0}) \rightarrow$ DTFT of nx(n), R and I are the real and imaginary parts of the Fourier transform, respectively (proof of Eq. 1 is highlighted in Appendix), and $|X_c(e^{j\omega_0})|$ is cepstral smoothing version of $|X(\omega_0)|$.

Algorithm 1 GDCC Features Extraction

1. $x(n)$ = short-time speech signal segment
2. y(n) = n.x(n)
3. $X(e^{j\omega_0})$=DTFT{x(n)}=$X_R(e^{j\omega_0})+jX_I(e^{j\omega_0})$, $Y(e^{j\omega_0})$=DTFT$\{nx(n)\} = Y_R(e^{j\omega_0}) + jY_I(e^{j\omega_0})$
4. $\tau(e^{j\omega_0}) = \frac{X_R(e^{j\omega_0})Y_R(e^{j\omega_0})+X_I(e^{j\omega_0})Y_I(e^{j\omega_0})}{|X_c(e^{j\omega_0})|^2}$
5. GDCC = DCT$\{\tau(e^{j\omega_0})\}$

The Group Delay Cepstral Coefficients (GDCC) features are derived from GDF. The computation of GDCC involves a series of steps (as shown in Algorithm 1) to capture the speech signal's phase information and spectral dynamics. Because of its capacity to capture fine-grained temporal and spectral properties, this feature has demonstrated promising results in a variety of speech processing applications [9]. For the extraction of GDCC features, the Hamming window was 25 *ms* duration, the shift was 10 *ms*, and the *20-D* (to be discussed shortly in Section 4.1) cepstral features were used for experiments. Figure 2 shows the functional block diagram of proposed GDCC feature based methodology.

2.1. Discussion

This Section discusses the significance of the selected feature vector. The concept of GDCC revolves around the concepts of spikes and zeros in the signals. The spikes in the signals are also often known as formants in speech technology. These formants in the z-plane of a real-time speech signal are in the pattern of the mixed-phase system, which means they lie inside and outside the unit circle. In minimum phase signal (like natural speech), if the roots are not near the unit circle, then GDF identifies the signal correctly [13]. In Figure 1 (a), Panel 1 and Panel 2 describe the waveforms of the signal with and without pop noise, respectively. Figure 1 (b) Panel 2 shows the enlarged pop noise region of (0.8 - 1.4 *s*) from (a), which shows the effect of a sudden closure of vocal folds resulting in pop noise formation which is decayed in Panel 1, as it is recorded from a far distance [4]. Figure 1 (c) signifies the difference between the magnitude spectrum of recorded and original speech. Figure 1 (d) represents full-frequency scale GDF, which is further magnified and analyzed in Figure 1 (e) as previous studies found that pop noise lies at a low-frequency region [14]. Figure 3 illustrates both the spectrogram and the group delay spectrogram of the audio signal. It is evident that the frequency bins in the group delay spectrogram exhibit higher resolution in lower frequency regions compared to the spectrogram. However, when

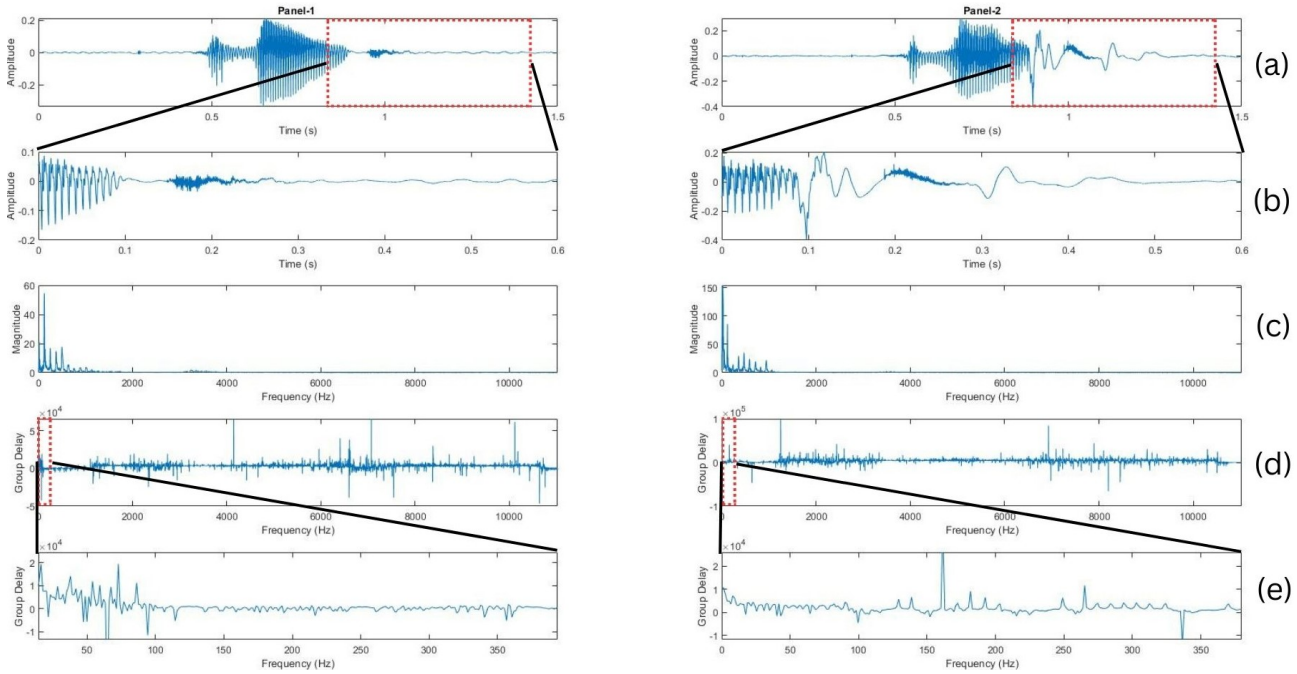

Figure 1: Panel-1 (signal without pop noise), and Panel-2 (signal with pop noise): (a) the original speech signal, (b), (c), (d), (e) shows the short-time signal, corresponding magnitude spectrum, group delay function, and its selected region for 0-400 Hz of the selected signal (as shown via the dotted box in Figure 1(d)), respectively.

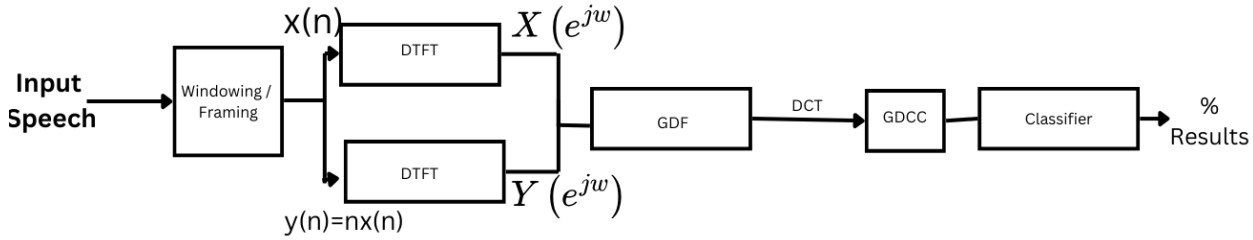

Figure 2: Functional block diagram of proposed GDCC-based VLD system.

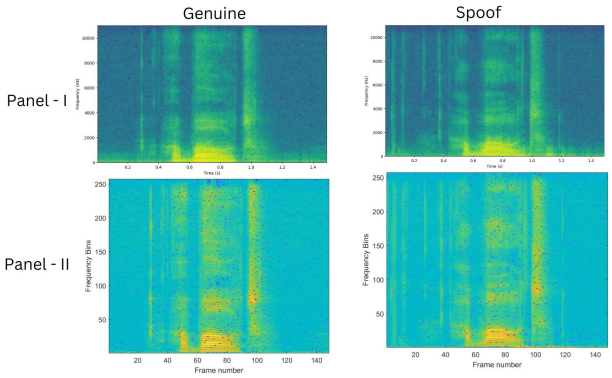

Figure 3: Panel-1 represents simple spectrogram, and Panel-2 represents group delay spectrogram.

examining the spectrogram, there is not a noticeable distinction between genuine and spoof signals. According to the authors' first knowledge, this may be the key reason why the GDCC feature vector outperforms another feature vector *w.r.t.* in its ability to capture unique phase-based characteristics, unlike other magnitude-based features.

3. Experimental Setup

3.1. Dataset Employed

In real-world cases, when a fraudster intends to carry out a Synthetic Spoofing Attack (SSA), they typically need to acquire the target subject's original audio. The most straightforward method to acquire this data involves recording the audio of the specific speaker of interest and subsequently utilizing that recorded audio to integrate a Speaker Specific Adaptation (SSA) into the Automatic Speaker Verification (ASV) system. The pop noise may not be captured since the collection of audio will be carried out from far away [3]. In such cases, it is feasible to identify and distinguish between real *vs.* spoofed speech by taking advantage of the lack of pop noise in replayed samples.

We have been using the POCO dataset for this model, which has been recently released [4]. A total of 66 people, 34 were female, and 32 were male, whose age range is from 18 to 61

years [4]. All the words are in one language (English), and 44 verbs were present in the dataset, sampled at 22050 Hz. The dataset has been released in 3 sub-parts, i.e., RC-A (microphone recordings), RP-A (eavesdropping), and RC-B (microphone array recordings). However, RC-B (a portion of data related to spoofed speech) was not executed during this study as it contains microphone array recordings. The dataset was split up

Table 1: Details of POCO Dataset. After [4].

	No. of Samples	**Total Subjects**	**Male Subjects**	**Female Subjects**
Training	13640	53	26	27
Evaluation	3432	13	6	7

into two categories: training (80%) and evaluation (20%) of all the utterances. Each of these subgroups is split equally between genuine and spoof audio samples. We also made sure that every speaker subset was unique, and that the proportion of male-to-female speakers was kept constant. Table I displays the data distribution statistics in the training and evaluation subgroup. All the detailed information about the dataset is given in [4].

3.2. Other Spectral Features Used

We employed three commonly used features as comparison with baseline features, i.e., Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Gammatone Frequency Cepstral Coefficients (GFCC) for comparison in this study. Motivated from ability of MFCC for human perspective of speech evaluation on mel scale, and LFCC for human perspective of speech evaluation on linear scale. On the other hand, GFCC is renounced to capture spectral characteristics of human audios, makes it perfect for comparison alongside MFCC, and LFCC.

3.3. Classifiers Used

We have performed the experiments using CNN [15], CRNN [16], and LCNN [17] as classifiers [18]. CNN and CRNN are employed as binary classifiers, where the two classes are speech with pop noise (i.e., genuine human voice) and without pop noise (i.e., spoofed voice) [19]. A learning rate of 0.001 is used in the construction of each model using the Adam optimizer. The loss function for binary classification jobs is binary cross-entropy, with a batch size of 64. The experiment employs *5*-fold cross-validation, with 100 epochs executed during training.

4. Experimental Results

4.1. Effect of Dimension of GDCC feature vector

Before performing comparison experiments, it is important to optimize parameters for a particular feature vector. In contrast to another feature set, GDCC offers the advantage of having only a few parameters that require optimization, with one of them being the dimension of the feature vector. Table 2 represents the effect on accuracy *w.r.t.* change in dimension of the feature vector, which shows state that as we increase the dimension of the feature vector, the accuracy increases to a certain extent, after-by the accuracy starts decreasing due to possible reduction in features as the dimensions are higher than required (*20-D*). This study focused on detecting pop noise through a

Table 2: Effects of Dimension of GDCC Feature Vector using CNN Classifier

Dimension	**Accuracy (in %)**	**F1-Score**	**EER**
10	76.43	75.82	23.77
13	74.07	74.33	26.75
16	78.06	77.68	22.14
20	**79.32**	**79.54**	**20.52**
23	78.53	78.64	21.21
26	78.09	77.96	23.02
30	77.51	78.08	22.14

binary classification task. The dataset was labeled "1" for pop noise files, and "0" for non-pop noise files. In Figure 4, we have compared wordwise accuracies for MFCC, LFCC, GFCC, and GDCC, all using the same CNN classifier. It can be observed from Fig. 4 that MFCC-CNN gives the least accuracy, while the accuracies of GFCC-CNN and LFCC-CNN are almost comparable. In some instances, GFCC-CNN performs better than LFCC-CNN or vice-versa; such behavior could be because some words contain low-intensity pop noises. However, we can see that GDCC-CNN outperforms all other feature sets.

Table 3: Accuracy (in %) for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	59.12	72.00	74.80
LFCC	76.74	72.73	77.07
GFCC	65.33	68.35	67.92
GDCC	**79.55**	**74.94**	**77.59**

Table 4: F1-score for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	0.5738	0.7100	0.730
LFCC	0.7357	0.7449	0.7451
GFCC	0.6478	0.6752	0.6861
GDCC	**0.7913**	**0.7432**	**0.7535**

Table 5: EER (in %) for corresponding features and classifiers

Feature Set	CNN	CRNN	LCNN
MFCC	43.32	23.69	26.51
LFCC	23.31	26.04	22.96
GFCC	33.88	32.36	32.79
GDCC	**20.68**	**23.54**	**22.26**

The experimental results in Table 3, Table 4, and Table 5 suggest that by using GDCC features in combination with CNN, the VLD system achieved relatively better performance. The accuracy score obtained was 79.55%, indicating that the system accurately classified the majority of the samples. The F1-score,

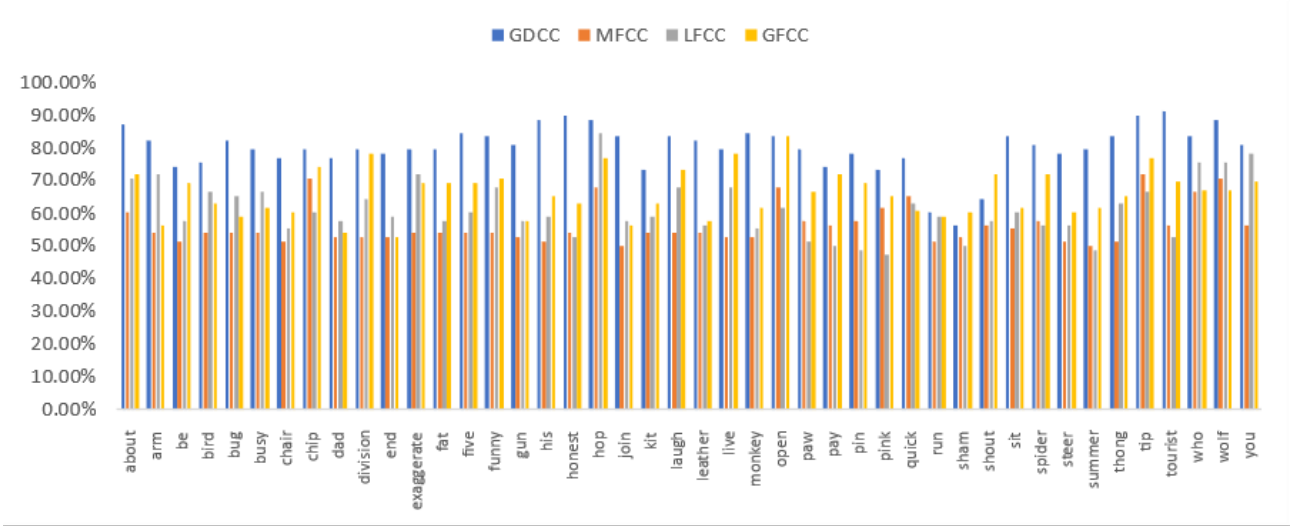

Figure 4: Comparision of wordwise accuracy (in %) for GDCC, MFCC, LFCC, and GFCC features using CNN classifier.

which stands at 0.7931, implies a good equilibrium between precision and recall. This suggests that the system managed to minimize false positives and negatives to a certain extent. The EER score of 20.68% demonstrates the system's ability to achieve a balanced error rate. These findings suggest that phase-based features capture fine-grained temporal and phase-related properties of speech signals, allowing for better discrimination between genuine human voices and spoof or prerecorded signals. Utilizing GDCC features in conjunction with deep learning models can bolster the security and reliability of voice biometric systems.

On the other hand, GDCC features are beating each feature set individually even after changing the classifier. GDCC-LCNN and LFCC-LCNN are close, but GDCC-LCNN still has a minute better result than LFCC-LCNN. Hence, GDCC-CNN is promising for pop noise detection in VLD, offering improved accuracy compared to the other feature extraction techniques.

4.2. Comparison With existing Approaches

In this subsection, we contrast our proposed methodology with established approaches. We achieved higher accuracy compared to recent optimal existing approaches, which utilize acoustic features derived solely from the magnitude spectrum of the speech signal, overlooking the phase spectrum of the speech signal. By comparing this work and showing the superiority of accuracy, we can say that phase-based characteristics carry a greater amount of information as compared to the magnitude spectrum of the pop noise signal. Table 6 shows that we obtained a total of 7.73% increase in accuracy as compared to the baseline work [20].

4.3. Evaluation of Latency Period

The latency period, a critical factor in numerous real-time applications, measures the time delay between the occurrence of an event and its detection [21]. Understanding the performance of different latency calculation techniques is vital for optimizing the efficiency and accuracy of such systems. The minimum amount of speech duration needed to obtain the highest possible classification accuracy is called *latency period*. Figure 5 shows the analysis of the latency period for 3 selective features, i.e., GDCC, MFCC, and LFCC.

Table 6: Comparison with existing works in existing literature.

Study	Feature	Frequency Range (in Hz)	Classifier	Accuracy
[20]	STFT	0-11025	CNN	71.81
[14]	Bump wavelet	0-40	CNN	74.69
(A)	Mel-STFT	0-40	CNN	76.39
Proposed	GDCC	0-11025	CNN	**79.54**

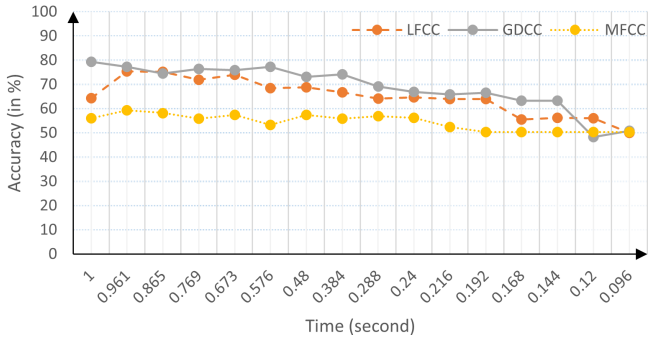

Figure 5: Analysis of latency period using accuracy (in %) as measure using CNN classifier

5. Summary and Conclusions

In this research, different classifiers were employed to assess multiple spectral feature sets, including MFCC, LFCC, GFCC, and GDCC. Despite altering the classifier, the experiments consistently indicated that GDCC exhibited higher accuracy compared to the other three spectral feature sets. These experimental results underscore the efficacy of GDCC features for detecting pop noise in the VLD task. These results are further

supported by a graph comparing wordwise accuracies, which reveals that MFCC-CNN is the least accurate while GFCC-CNN and LFCC-CNN perform similarly. The higher performance of GDCC-CNN raises the possibility of using it to enhance pop noise detection in VLD applications. GDCC-CNN offers improved accuracy compared to other feature extraction approaches, making it a potential option for pop noise identification in VLD. However, GDCC features are computed based on the short time window of speech signals. While they capture information about the spectral shape and group delay properties within each window function, they may lack the ability to capture long-term contextual information, such as prosody or semantic meaning. Also, they use sliding window analysis, resulting in a high-dimensional feature space. This can pose challenges in terms of computational complexity and storage requirements, especially when dealing with large datasets or real-time applications. We aim to focus our future research on exploring the latency period of GDCC features, with the goal of assessing their feasibility for deploying VLD systems in real-world scenarios.

6. References

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

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre 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] Shihono Mochizuki, Sayaka Shiota, and Hitoshi 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.

[4] 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*, 2020, pp. 1081–1085.

[5] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education India, 2002.

[6] 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.

[7] Mohit Dua, Rajesh Kumar Aggarwal, and Mantosh Biswas, "GFCC based discriminatively trained noise robust continuous asr system for hindi language," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 2301–2314, 2019.

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

[9] Hema A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, pp. 745–782, 2011.

[10] E.J. Hannan and P.J. Thomson, "Estimating group delay," *Biometrika*, vol. 60, no. 2, pp. 241–253, 1973.

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

[12] Alan V Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, pp. 458–465, 1969.

[13] Xiangwei Zhu, Yuanling Li, Shaowei Yong, and Zhaowen Zhuang, "A novel definition and measurement method of group delay and its application," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 1, pp. 229–233, 2008.

[14] Priyanka Gupta, Siddhant Gupta, and Hemant A. Patil, "Voice liveness detection using bump wavelet with CNN," in *9th International Conference on Pattern Recognition and Machine Intelligence*, 2021, Kolkata, India.

[15] Keiron O'Shea and Ryan Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015 {Last Accessed date : 20^{th} February, 2024}.

[16] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte, "Recurrent convolutional neural networks: a better model of biological object recognition," *Frontiers in Psychology*, vol. 8, pp. 1551, 2017.

[17] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs, "Locally scale-invariant convolutional neural networks," *arXiv preprint arXiv:1412.5104*, 2014, {Last Accessed : 9^{th} February, 2024}.

[18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[19] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[20] Siddhant Gupta, Kuldeep Khoria, Ankur T. Patil, and Hemant 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)*. IEEE, 2021, Tokyo, Japan., pp. 775–779.

[21] James William Topliss, Victor Zappi, Andrew McPherson, et al., "Latency performance for real-time audio on beaglebone black," 2014.