Circuits, Systems, and Signal Processing (2022) 41:3811–3831
https://doi.org/10.1007/s00034-022-01957-0

Generative and Discriminative Modelling of Linear Energy Sub-bands for Spoof Detection in Speaker Verification Systems

Suvidha Rupesh Kumar[1] · **B. Bharathi**[1]

Received: 5 April 2021 / Revised: 24 December 2021 / Accepted: 28 December 2021 /
Published online: 20 January 2022

Abstract

Classification of genuine and spoofed utterance is the basis for most of the countermeasure detecting spoof attacks on automatic speaker verification system. The choice of a good discriminating feature and a complementing classifier adds to the robustness of the countermeasure. Cepstral coefficients of the linear sub-band energy analysis have proved its worth in countering unknown attacks as witnessed by the literature. The intention behind the proposed work is to assess the behaviour of a spoof detection countermeasure using linear frequency cepstral coefficients with both generative and discriminative classifiers. The same are considered as baseline systems for further analysis. Parallelly, the paper proposes modifications to the traditional weighting function used in the retrieval of energy sub-bands on linear scale in order to leverage its full potential in spoof detection. The weighting function used is Gaussian, and hence, the modified feature is referred as GaussFCC. The aforementioned analysis is carried out on non-pre-emphasised utterances. The classifiers used are Gaussian mixture model (generative) and bidirectional long short-term memory (discriminative) classifiers. The empirical results show that the generative classifier has performed significantly in the detection of spoof attacks under logical access condition and discriminative classifier has shown drastic improvement in spoof detection under physical access condition over the generative model. Tandem detection cost function for logical access scenario (LA) using GMM classifier is 0.000 for development data and 0.113 for evaluation data, and in physical access scenario using BiLSTM classifier, it is 0.030 for development data and 0.044 for evaluation data. A detailed comparative analysis of the performance of the countermeasure is carried out based on different types of attacks, features, classifiers and utterances from female and male speakers.

✉ Suvidha Rupesh Kumar
suvidhark@ssn.edu.in

B. Bharathi
bharathib@ssn.edu.in

[1] Department of CSE, SSN College of Engineering, Chennai, Tamil Nadu, India

Keywords Gaussian filter · Spoof attack · Automatic speaker verification system · Gaussian mixture model · ASVspoof 2019 · Speech utterance · GaussFCC · GMM · BiLSTM

1 Introduction

The literature in spoof detection on automatic speaker verification system has proof of vulnerability of the system to different spoof attacks [7] and the development of countermeasures [31] for the same. The researchers have explored the possibility of improving the robustness of the countermeasure through feature enhancement approaches [4,16,35] and varying the choice of the classifier [37][15].

With the advent of the biometric systems [8], the area of identification and verification of the biometric traits [23] has gained momentum. The concern is about the reliability of the system, as most of these applications are used for authentication purposes. One of the biometric traits used is the voice print [21]. An automatic speaker verification system, on the authentication front, should prevent any fraudulent entry. All kinds of biometric systems are vulnerable to spoof attacks including automatic speaker verification (ASV) system [7]. The detection of spoof attacks on the speaker verification system [12] is and will be an open research with the emergence of high quality techniques imitating the naturalness of human voice [17,25] posing high threat to the system. The research community has contributed towards this end with feature re-engineering [22] and classification of these features with various classifiers [9,18,37].

Linear frequency cepstral coefficient (LFCC) features-based countermeasure has performed significantly in the detection of unknown attacks as analysed in the literature. Many researchers have chosen LFCC as the optimal candidate to fuse with other features to improvise the performance of a countermeasure system in spoof detection. This has been the contributing feature under both LA and PA scenario catering to the generalised countermeasure countering different kinds of spoof attacks on ASV system. This has motivated the authors of the paper to leverage the potential of LFCC through fine tuning the weighting factors in the linear sub-band aiming to boost up the discriminating characteristic of cepstrum and suppress noise as well. The feature is tested on traditional generative classifier GMM and the discriminative classifier BiLSTM, known for its learning capability of long-term dependencies between sequence data which proved to be a complementary classifier for GaussFCC.

In their previous work [16], the authors have focused on capturing most of the significant energy variations into few cepstral coefficients by obtaining the energy variation pattern forming the basis for FBCC (filter-based cepstral coefficient) feature extraction. This showed a significant improvement of countermeasure under both LA and PA condition. In the current paper, the authors have chosen BiLSTM as a classifier to classify the genuine and spoofed utterances using GaussFCC features. Here the cepstral coefficients capture the variations at feature level and the BiLSTM classifier learns bidirectional long-term dependencies between sequence cepstral coefficients for classification. Through this, a significant performance improvement of the countermeasure could be achieved under physical access condition and comparatively good perfor-

mance under logical access condition. Both the works indicate that a complementing feature-classifier combination would do better spoof detection by well-capturing time-varying information within the sequence adding to the discrimination of spoof from genuine utterance.

Organisation of the rest of the paper is as follows: Sect. 2 discusses about LFCC in unknown spoof detection on ASV systems. Section 3 discusses about the corpus and classifiers used. Section 4 explains the proposed feature for spoof detection. Section 5 discusses the performance analysis of the countermeasure using the proposed feature with GMM and BiLSTM classifiers. This is followed by Sect. 6 for conclusion and Sect. 7 for acknowledgement.

2 Performance of LFCC in Unknown Spoof Detection on ASV Systems

Enhancement of the discriminating nature of feature involved in speech detection is indispensable with the increasing naturalness of the synthetic speech and replayed speech. A feature required for classification problem needs to meticulously capture the inter-class discrimination or intra-class affinity among the data under classification. Challenge related to the selection of such features varies based on the application at hand. The research community is confronted with the task of spoof detection on the speaker verification system. This task demands a robust feature which could capture the traces of a spoof attack on the utterance presented to the system. Research works have led us to countermeasures with the robust features in detecting such attacks. The countermeasure either differs in feature used in the front-end processing or classifier used in back-end classification with good performance.

The feature re-engineering has witnessed the modified versions of the filterbank to emphasis the frequency component of interest [27,36]. The traditionally used features those that fall in this category are linear frequency cepstral coefficients (LFCCs), mel frequency cepstral coefficients (MFCCs) and inverted mel frequency cepstral coefficients (IMFCCs) [33]. Most of these features differ in the sub-band analysis depending on the application under consideration. For example, LFCC, MFCC and IMFCC are based on linear scaled, mel-scaled and inverted mel-scaled sub-band analysis, respectively.

Previous research [24] has dealt with a detailed comparative analysis of features which proved to enhance the performance of countermeasure in detecting spoof attacks on ASV system under LA condition. The performance analysis of countermeasure with these features is experimented with ASV spoof 2015 dataset [32]. ASV spoof 2015 dataset deals with spoofed utterances generated from ten different algorithms (S1–S10), and the evaluation set contains both known (S1–S5) and unknown (S6–S10) attacks. The features analysed under short-term power spectrum are filterbank-based cepstral features, namely RFCC, LFCC, MFCC and IMFCC, all-pole modelling-based cepstral features, namely linear prediction cepstral coefficients (LPCC), and perceptual linear prediction cepstral coefficients (PLPCC), spectral flux-based feature, namely sub-band spectral flux coefficient (SSFC), sub-band spectral centroid-based features, namely sub-band centroid frequency coefficients (SCFC), and spectral centroid magnitude coefficients (SCMC) and under short-term phase features are modified

group delay function (MGDF), all-pole group delay function (APGDF), cosine-phase function (CosPhase) and relative phase shift (RPS) and under spectral features with long-term processing are modulation spectrum (ModSpec), shifted delta coefficients (SDCs), frequency domain linear prediction (FDLP) and mean Hilbert envelope coefficients (MHECs). These 17 features are experimented with GMM and SVM classifiers. The $\Delta\Delta$ value of LFCC proved to be significant for unknown attacks with an average equal error rate (EER%) of 1.67 using GMM classifier outperforming the other features.

In [33], the authors have presented a comparison of LFCC, MFCC, IMFCC and CQCC features under PA condition. The experiments were conducted on two datasets, namely ASVspoof 2017 and BTAS 2016. The EER% of LFCC for ASVspoof 2017 dataset was 3.31 and 2.04 for unknown attacks.

In [28], the performance of LFCC for unknown attacks is listed with EER% of 1.670 on ASVspoof 2015 dataset. It is shown to outperform cepstral coefficients and change in the instantaneous frequency (CFCC-IF) feature, system with i-vectors based on MFCCs, mel frequency principal coefficients and cosine-phase principal coefficients feature and magnitude- and phase-based feature.

In [11], with score-level fusion of LFCC and TECC (Teager energy cepstral coefficients) on ASVspoof 2017 dataset, the countermeasure outperformed fusion of TECC with MFCC and CQCC as well.

Hence, the LFCC feature set has proved to be consistently good for detection of a spoof attack on ASV systems as a stand-alone and as a good candidate feature for fusion as well.

3 Database and Classifiers

3.1 Speech Corpus with Spoof and Bonafide Utterances

Speech Corpus used for the proposed work is ASVspoof 2019 corpus [29] detailed in Table 1. The dataset is categorised into logical access (LA) and physical access (PA) scenarios. LA access condition consists of speech synthesised and voice converted utterances. PA access condition consists of recorded and replayed utterances. Each of LA and PA consists of bonafide and spoofed utterances for training, development and evaluation. The number of speakers are 8 male and 12 female. The duration of utterances in the dataset is in the range of 1–11 s.

The training and development sets contain known attacks, and evaluation set contains 2 known and 11 unknown spoofing attacks. There are six known attacks, of which two are voice conversion (VC) systems and four from text-to-speech synthesis (TTS) system. TTS systems use either waveform concatenation or neural network-based speech synthesis using a conventional source-filter vocoder or a WaveNet-based vocoder. Among 11 unknown systems, there are two VC, six TTS and three hybrid TTS-VC systems. These are implemented with various waveform generation methods including classical vocoding, Griffin-Lim, generative adversarial networks, neural waveform models, waveform concatenation, waveform filtering, spectral filtering and

Table 1 ASVspoof 2019 speech corpus

Subset	#utterances			
	Logical access (LA)		Physical access (PA)	
	Bonafide	Spoof	Bonafide	Spoof
Training	2580	22,800	5400	48,600
Development	2548	22,296	5400	24,300
Evaluation	7355	63,882	18,090	11,6640

their combination. The references related to the known and unknown attacks and their implementation details are mentioned in [29].

3.2 Generative and Discriminative Classifiers

The spoof detection using GaussFCC is explored on GMM and BiLSTM. Though both the systems have shown significant improvement over the LFCC-based baseline system under both LA and PA conditions, the one with GMM classifier has performed well under LA condition and the latter has performed well under PA condition comparatively. In [1], the authors present a detailed discussion on generative and discriminative models.

3.2.1 GMM Classifier

Gaussian mixture model is a generative approach where the joint distribution is considered in the model. Traditional Gaussian mixture model (GMM) [6] is used here. Two such models are generated, one for GaussFCC features from spoofed utterances and a second one for GaussFCC features from bonafide utterances.

The GaussFCC features from the test utterance are extracted and are presented to the spoofed and bonafide GMM. The log likelihood scores S_b and S_{sf} are computed for the bonafide and spoofed model, respectively. The log likelihood difference is computed as $\lambda = S_b - S_{sf}$. Here, λ is the final score of each of the test utterance. The positive value of λ would classify the utterance as bonafide and spoofed otherwise. The GMM classifier used is shown in Fig. 1.

3.2.2 BiLSTM Classifier

The problem of limited long memory capability of RNN is addressed by LSTM units with the concept of an additional hidden state to h(t), the cell state C(t). Gates remove or add information to C(t) based on the input value x(t) and the hidden value h(t-1). Gates are implemented using sigmoidal layer. The feature that makes LSTM more appealing in the field of speech processing is its "long-term dependencies" [10]. Hence, the bidirectional LSTM has the property of "long-term dependencies".

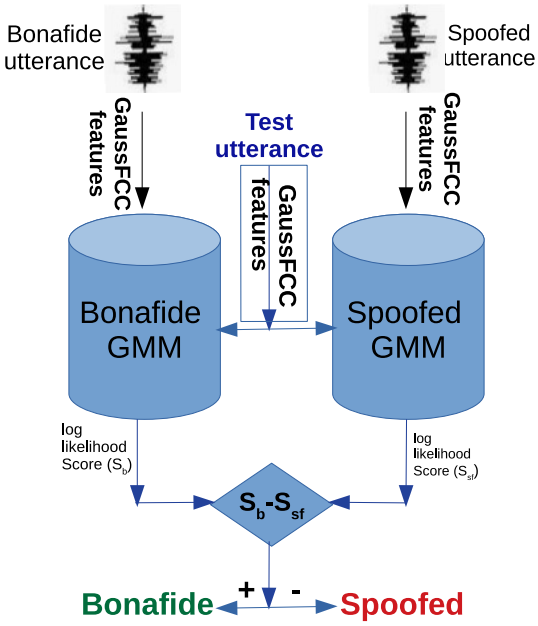

Fig. 1 Spoof detection in ASV systems using GMM classifier

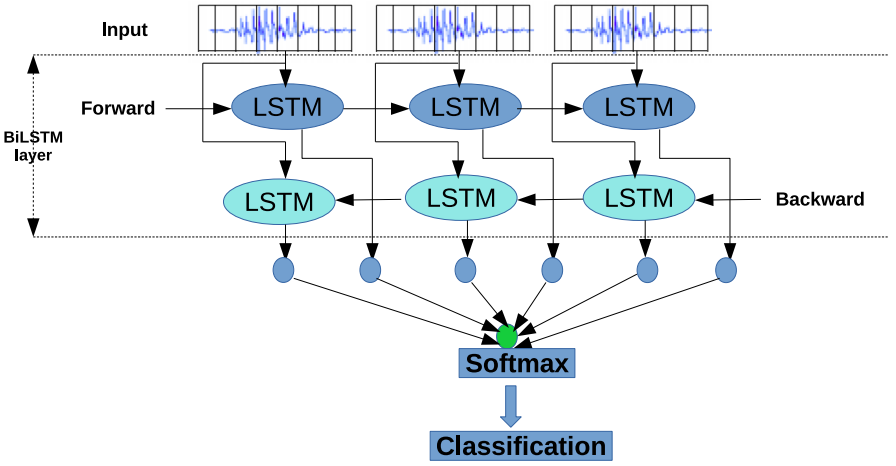

Fig. 2 Spoof detection in ASV systems using BiLSTM

BiLSTM is a bidirectional LSTM, in which the signal propagates in both backward and forward direction in time. The simple architecture of the BiLSTM classifier used here is shown in Fig. 2.

The number of frames generated for each utterance would differ based on the duration of the speech. But for each frame, the number of cepstral coefficients retrieved would remain the same as 120 including dynamic coefficients, namely delta and delta–delta coefficients. The padding could be reduced by sorting the training and testing data by sequence length, and choosing a mini-batch size so that sequences in a mini-

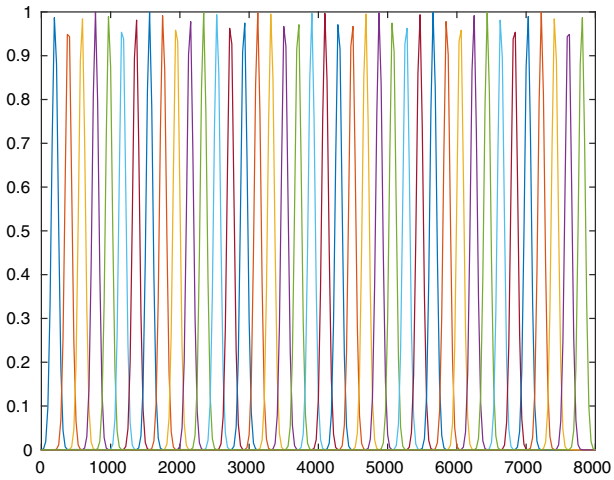

Fig. 3 GaussFCC filterbank

batch have almost similar length. This is an added advantage when the application deals with speech utterances [14].

4 Proposed Feature: GaussFCC

In our current research work, we propose to modify the weighting function of the linearly scaled LFCC feature to enhance its capability in spoof detection on ASV systems. Owing to the usage of Gaussian weighting energy sub-band for the retrieval of cepstral coefficients, the proposed feature is referred as GaussFCC. The Gaussian filter is formulated using the Gaussian membership function represented as Gaussian(x:c,s), where c, s represent the mean and standard deviation, respectively. The filterbank is shown in Fig. 3.

The number of filters used for the experiments are 40. The idea is to obtain 40 cepstral coefficients. The finer spectral details are captured by the higher-order cepstral coefficients [19], and hence, all the 40 cepstral coefficients are retained without discarding any. The experiments are conducted with energy sub-bands of GaussFCC closely resembling that of LFCC. This is to experiment with GaussFCC performance when sub-bands closely resemble that of LFCC. There are two sets of experiment investigated in this paper based on the methods used to compute the standard deviation (σ). In the first set of experiment, the σ is approximated to a value tuned using α-factor. The performance of countermeasure is found to be good under LA condition when σ is computed with α set to 3 and to 2 under PA condition in the given equation,

$$\sigma = \frac{(f_{i+2} - f_i)}{2 * \alpha} \tag{1}$$

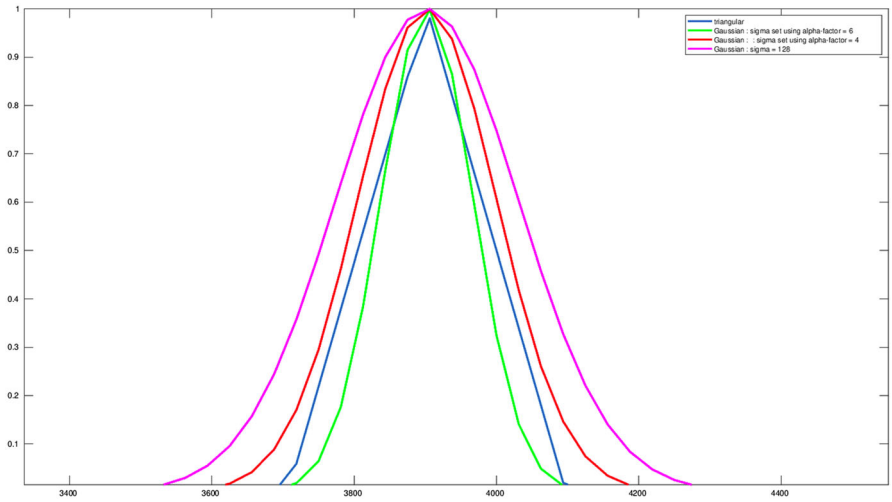

Fig. 4 GaussFCC filter with different σ values along with triangular filter (refer Eq. 1)

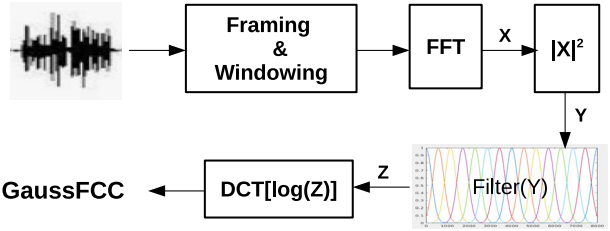

Fig. 5 GaussFCC computation flow

where f_i and f_{i+2} are the lower and higher values of bandwidth (BW) and BW is $\frac{(f_{i+2}-f_i)}{2}$. The second set of experiments is tested with a constant value σ=128 at the neighbourhood of full width half maximum region yielding good results. There is a folded Gaussian filter set at the minimum and maximum frequency in this case. The results of the aforementioned two sets of experiments are discussed under result analysis section. A filter showing the placement and bandwidth of the Gaussian filter used for the experiments is shown in Fig. 4 along with the triangular filter at the same position for comparison, when the number of filters is 40 over a frequency ranging from 0 to 8000Hz.

The GaussFCC feature extraction stages are shown in Fig. 5. The pre-emphasis of the speech utterance is not performed throughout the experiments.

The additional information captured by GaussFCC as compared to the one used in LFCC is shown in Fig. 6. The filtered energy captured for a spoofed utterance that is detected by GaussFCC feature and missed by LFCC feature is depicted in Fig. 6. In Fig. 7, first, second, third, fourth, twentieth and fortieth cepstral coefficients obtained from a randomly selected sequence of frames are shown. A positive value of cepstral coefficient indicates that spectral energy is more concentrated in the low frequency

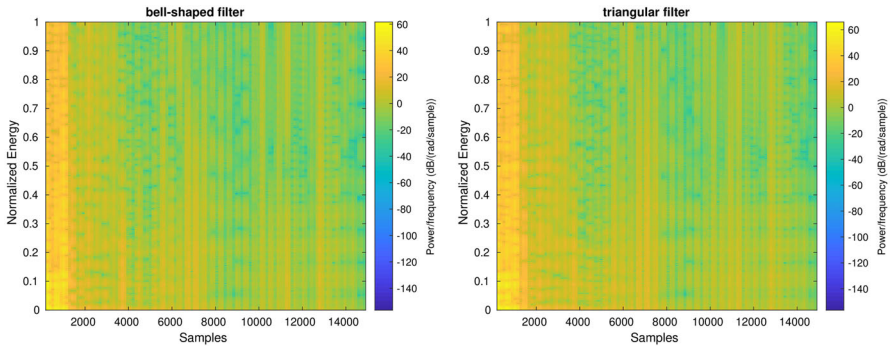

Fig. 6 Visualisation of the energy captured by GaussFCC (left) and LFCC (right) features (the energy captured is of spoofed utterance detected by GaussFCC and missed by LFCC)

region and a negative value indicates that most of the spectral energy is concentrated on the high frequency region [5]. The highlighted values in the figure show the significant variation that has taken place with the change in weighting function while processing the same utterance. As mentioned earlier, the idea of suggesting the change in weighting function is to use the linearly scaled sub-band analysis to its full potential following its success as witnessed in the literature. The statistical significance of the GaussFCC approach is studied through the entropy computation of the probability distribution obtained from the histograms of utterances after the application of triangular and Gaussian filters. The entropy equation used is $E = -\sum_{i=1}^{n}[P_i log_b(P_i)]$ and is the one introduced by Shannon [26]. The entropy of the probability distribution obtained from the triangular(E_{tri}) and Gaussian(E_{gauss}) weighted filtered energy under LA condition is 0.3571 (E_{tri}) and 0.2930 (E_{gauss}) and under PA condition is 0.4026 (E_{tri}) and 0.3353 (E_{gauss}). The entropy values indicate that Gaussian weighting tends to elicit information more than the triangular weighting function [3]. The robustness of the feature is further evident through the empirical results obtained when experiment is conducted with ASVspoof 2019 corpus which is discussed under the result analysis section. As far as linear filters are concerned, the energy analysis bands are linearly scaled. The outcome justifies the intuitive idea that the linear filter captures sufficient information required to detect traces of spoof attack.

4.1 Pre-emphasis or No Pre-emphasis in Spoof Detection

The paper analyses the performance of the countermeasure over feature extracted from non-pre-emphasised utterance. Figure 8 shows the spectrum of an utterance before and after pre-emphasis. In Fig. 8, utterance U1 is identified as spoof even after pre-emphasis, but utterance U2 is not identified as a spoof after pre-emphasis. The spectrum shows the energy suppressed due to pre-emphasis in Fig. 8c and d. There is difference observed in a pre-emphasised signal from the original one, when the sound is played. The pre-emphasis filter used for analysis is $H(z) = 1 - 0.97z$. U1 and U2 are the utterances captured under PA scenario. The noise could be a trace of the attack from the record and replay devices or the ambience.

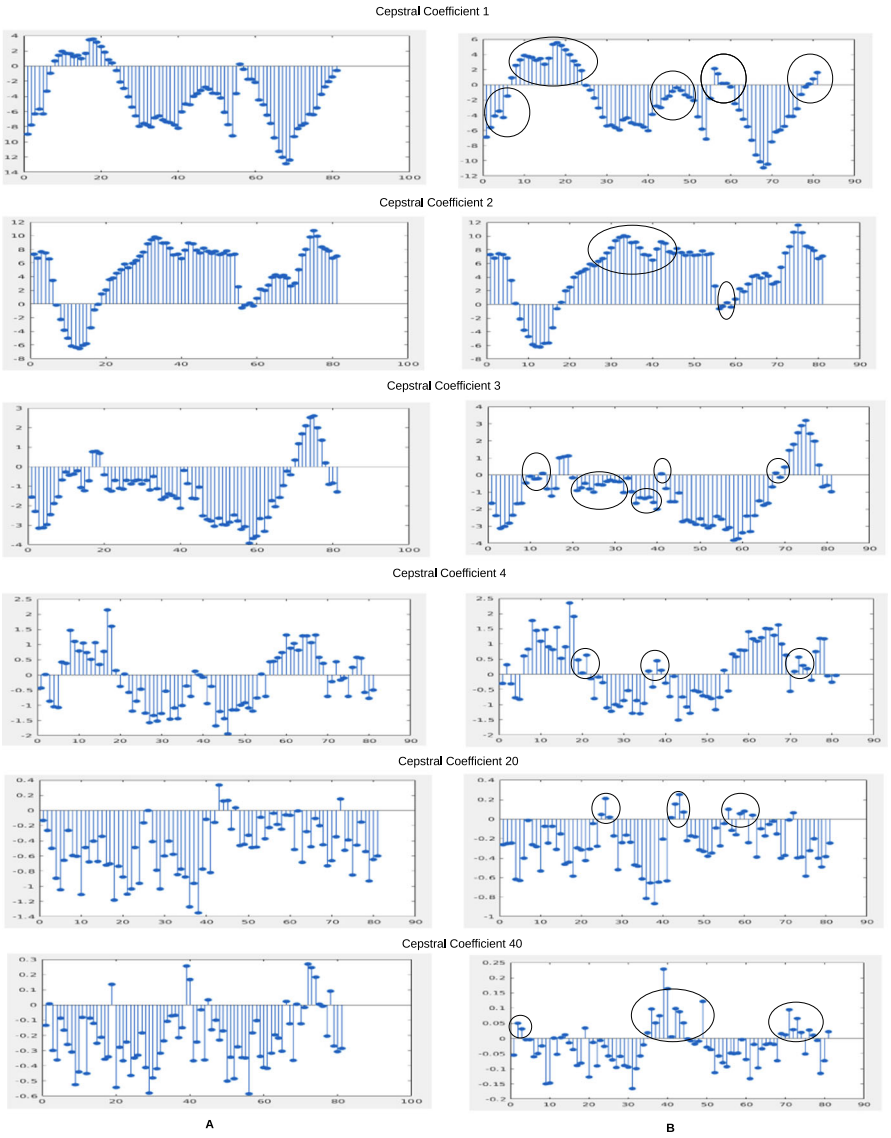

Fig. 7 Cepstral coefficients obtained in LFCC (**A**) and GaussFCC (**B**) features (significant variations are highlighted using circles)

5 Performance Analysis

5.1 Experimental Setup

The details of the corpus are shown in Sect. 3. The metric used for performance analysis is a minimum tandem detection cost function (min t-DCF) [13]. The utterance is not

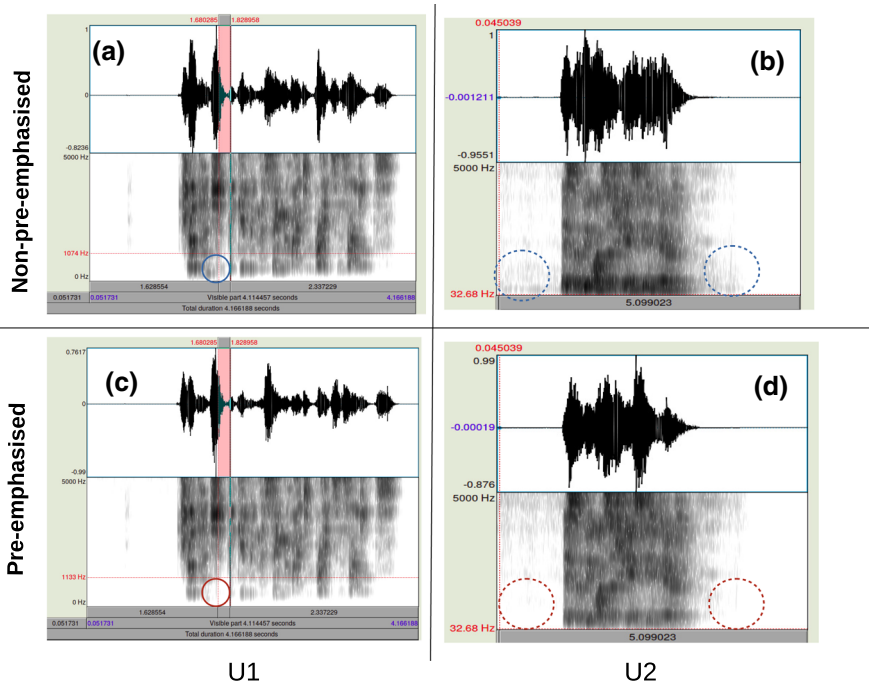

Fig. 8 Spectrum analysis of speech utterance from ASVspoof 2019 dataset before and after pre-emphasis. **a**, **b** are non-pre-emphasised speech utterance; **c**, **d** are pre-emphasised speech utterance

pre-emphasised. Framing is performed on the signal with a frame length of 20 ms with an overlap of 10 ms. Windowing function used is Hanning window. 40 Gaussian filters are used for sub-band analysis of energy. 40 cepstral coefficients are retrieved. Considering the static and dynamic values, the feature set consists of 120 coefficients. For the GMM classifier, the number of mixture components considered for LA is 512 and for PA is 256. Experiments were performed on development and evaluation data under both LA and PA scenarios for 100 iterations each. BiLSTM classifier used for classification consists of hidden layer of 64 nodes. The input sequence size is 120. Finally a fully connected layer of two nodes, one for each bonafide and spoof class. The minibatch size is set to be 100. The experiment was repeated for 100 epochs. The gradient threshold is set to 1. In conventional backpropagation algorithms, error flowing backward tends to explode or vanish depending on the weights and this in turn hampers the learning of long time lags by the network. Gradient threshold is set to a constant to overcome this problem which is 1 here.

5.2 Analysis I (GaussFCC1): GaussFCC with σ Controlled Using α-factor

The performance of the countermeasure is studied with the filter bandwidth set using Eq. 1. The performance of the countermeasure was found to be significant using $\alpha = 3$

Table 2 Performance of two versions of GaussFCC using GMM classifier and BiLSTM classifier with no pre-emphasis

Classifier (feature)	Development data		Evaluation data	
	EER%	min t-DCF	EER%	min t-DCF
LA condition				
GMM (Baseline)	0.008970	0.000179	4.475076	0.116253
GMM (GaussFCC1)	0.000000	**0.000000**	4.608803	**0.113991**
GMM (GaussFCC2)	0.000000	**0.000000**	4.622646	**0.112665**
BiLSTM (Baseline)	0.112691	0.002967	7.081146	0.145365
BiLSTM (GaussFCC1)	0.112691	**0.002021**	9.296747	0.169124
BiLSTM (GaussFCC2)	0.002243	**0.000045**	6.665340	**0.098200**
PA condition				
GMM (Baseline)	10.705761	0.227850	12.531754	0.295905
GMM (GaussFCC1)	8.111111	**0.179483**	9.674032	**0.232519**
GMM (GaussFCC2)	9.278807	**0.200293**	11.298964	**0.276820**
BiLSTM (Baseline)	1.422840	0.042208	2.034369	0.059586
BiLSTM (GaussFCC1)	1.089506	**0.030206**	1.476146	**0.043980**
BiLSTM (GaussFCC2)	1.074074	**0.030361**	1.713737	**0.0508**

The highlighted values are the minimum of min t-DCF value compared to the baseline and the boxed values are that of the baseline

for attacks under the logical access condition. The performance improved for attacks under physical access condition when the α value was set to 2 in Eq. 1.

5.3 Analysis II (GaussFCC2): GaussFCC with σ Set to a Constant

The constant value for σ was chosen experimentally to be 128 while analysing the bandwidth at the neighbourhood of the full width half maximum region. Hence, the experiments and results analysis substantiate well the robustness of the countermeasure to counter the attacks under LA and PA scenario with our proposed feature GaussFCC. The performance of the countermeasure with two versions of the proposed feature GaussFCC1 and GaussFCC2 as discussed above is shown in Table 2 for attacks under LA scenario and under PA scenario, respectively. The experimental result shows the performance of the countermeasure using GMM and BiLSTM classifier as well.

5.4 Analysis III: Performance Analysis of Countermeasure on Individual Attacks Under LA and PA Conditions Using Generative and Discriminative Classifiers for GaussFCC1 and GaussFCC2 Features

In Table 3, the spoof detection rate of the countermeasure for the individual attacks under logical access condition is shown. The highlighted values are that of the best results within the classifier level as compared to its baseline system, albeit GMM

Table 3 Performance of countermeasure using GaussFCC on individual attacks under LA condition

Attack type	GMM classifier						BiLSTM classifier					
	LFCC (Baseline)		GaussFCC1 (sigma = BW/6)		GaussFCC2 (sigma = 128)		LFCC		GaussFCC1 (sigma = BW/6)		GaussFCC2 (sigma = 128)	
	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF
LA (Development data)												
A01	0.0000	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	0.0331	0.0009	0.0331	**0.0008**	0.0000	**0.0000**
A02	0.0000	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	0.0796	0.0018	0.0000	**0.0000**	0.0000	**0.0000**
A03	0.0000	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	0.0331	0.0009	0.0331	**0.0003**	0.0000	**0.0000**
A04	0.0331	0.0008	0.0000	**0.0000**	0.0000	**0.0000**	0.1127	0.0032	0.1592	0.0036	0.0331	**0.0003**
A05	0.0000	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	0.1592	0.0049	0.1127	**0.0039**	0.0000	**0.0000**
A06	0.0331	0.0003	0.0000	**0.0000**	0.0000	**0.0000**	0.1261	0.0038	0.0000	**0.0000**	0.0000	**0.0000**
LA (Evaluation data)												
A07	0.0000	0.0000	0.0000	**0.0000**	0.0000	**0.0000**	0.0985	0.0026	0.0577	**0.0015**	0.0407	**0.0013**
A08	0.0000	0.0000	0.0000	**0.0000**	0.0170	0.0003	0.2682	0.0081	0.2682	0.0091	0.3497	**0.0119**
A09	0.0000	0.0000	0.0000	**0.0000**	0.0170	0.0003	0.1392	0.0041	0.2682	0.0081	0.1867	0.0064
A10	14.3046	0.4258	12.8378	**0.3895**	15.0211	0.4304	0.3260	0.0101	0.7096	0.0220	0.3837	0.0124
A11	0.0815	0.0024	0.0815	**0.0018**	0.0577	**0.0018**	0.3022	0.0094	0.6519	0.0209	0.2852	0.0097
A12	3.1135	0.0808	3.1781	**0.0727**	3.9861	0.0973	0.3667	0.0116	1.2631	0.0427	0.4889	0.0162
A13	4.2951	0.1300	5.3545	0.1602	5.9249	0.1741	0.6757	0.0226	1.8505	0.0619	0.4889	**0.0162**
A14	0.5059	0.0131	0.4482	**0.0107**	0.8387	0.0225	0.5874	0.0193	1.7282	0.0515	0.3497	**0.0118**
A15	5.5752	0.1615	5.1677	**0.1516**	3.2358	**0.0958**	0.5534	0.0182	1.3038	0.0428	0.4482	**0.0139**
A16	0.0000	0.0000	0.0000	**0.0000**	0.0170	0.0003	0.1460	0.0049	0.1630	**0.0048**	0.0815	**0.0025**
A17	7.9383	0.2123	9.1776	0.2528	8.6887	0.2282	37.6647	0.9155	40.3708	0.9832	39.2775	**0.9514**
A18	0.0985	0.0032	0.1222	0.0041	0.1800	0.0044	10.7836	0.3174	26.1307	0.5588	5.2492	**0.1186**
A19	0.1392	0.0028	0.1392	**0.0024**	0.0407	**0.0009**	2.4379	0.0559	0.2614	**0.0078**	1.1646	0.0294

Table 4 Information on attack types under PA condition

	A	B	C
Attacker-to-talker distance (AtoT)	10–50	50–100	> 100
Replay device quality (RdQ)	Perfect	High	Low

Table 5 Performance of countermeasure in min t-DCF on varying both, the attacker-to-talker distance and the quality of device

AtoT/ RdQ	Development data (min t-DCF)			Evaluation data (min t-DCF)		
	Perfect	High	Low	Perfect	High	Low
GMM classifier (Baseline)						
10–50	0.632662	0.100499	0.047410	0.672573	0.262827	0.118619
50–100	0.406737	0.067983	0.043795	0.496077	0.174977	0.087037
> 100	0.344977	0.057432	0.026021	0.429901	0.145975	0.082706
GMM classifier (GaussFCC1)						
10–50	0.505180	0.067145	0.030873	0.538532	0.171540	0.0804054
50–100	0.322573	0.042196	0.028988	0.387406	0.128287	0.060159
> 100	0.275352	0.038956	0.021354	0.348011	0.107747	0.061419
GMM classifier (GaussFCC2)						
10–50	0.587551	0.095098	0.043811	0.640236	0.234680	0.111382
50–100	0.359282	0.062935	0.042188	0.438416	0.164727	0.080514
> 100	0.290874	0.057599	0.029060	0.384840	0.137871	0.078276
BiLSTM classifier (Baseline)						
10–50	0.069284	0.023194	0.014317	0.081791	0.031554	0.047080
50–100	0.054593	0.021083	0.021317	0.078274	0.035401	0.054572
> 100	0.056967	0.025675	0.021767	0.074529	0.047159	0.051098
BiLSTM classifier (GaussFCC1)						
10–50	0.052888	0.020775	**0.014957**	**0.057537**	**0.031206**	**0.033949**
50–100	0.037016	**0.016408**	**0.015064**	0.053410	**0.032787**	**0.037089**
> 100	0.041217	0.024436	0.019065	0.052918	**0.045190**	**0.037518**
BiLSTM classifier (GaussFCC2)						
10–50	**0.044809**	**0.019936**	0.017617	0.059490	0.045699	0.037038
50–100	**0.033685**	0.016970	0.016620	**0.051594**	0.047391	0.040982
> 100	**0.036634**	**0.022550**	**0.017973**	**0.048330**	0.060624	0.047402

Comparatively good performance obtained by BiLSTM between GaussFCC1 and GaussFCC2 features

classifier shows good spoof detection over BiLSTM classifier when the values are compared between the classifiers for each of the features. Hence, the generative classifier performs better than the discriminative classifier for spoof detection under LA condition.

Table 6 Number of male and female utterances in development and evaluation dataset of ASVspoof2019 corpus

Subset	#utterances							
	Logical access (LA)				Physical access (PA)			
	Bonafide		Spoof		Bonafide		Spoof	
	Female	Male	Female	Male	Female	Male	Female	Male
Development	1680	868	14,904	7392	3240	2160	14,580	9720
Evaluation	5072	2283	44,226	19,656	9990	8100	65,610	51,030

In Table 4, the information on the attacker-to-talker distance [34] and three categories of quality of the replay device used to record and replay the utterances as in ASVspoof 2019 corpus [2,20,30] is shown.

Based on the above information, the performance of the countermeasure is captured for the individual attack type under PA condition and is depicted in Table 5. The results show that discriminative (BiLSTM) classifier is good at detecting spoof under PA condition compared to the generative classifier. The comparison of values of GaussFCC1 and GaussFCC2 using BiLSTM classifier depicts that both the features are complementary and could be the optimal candidates for fusion at score level.

Table 5 shows that when the quality of the device is perfect and the attacker-to-talker distance is varied, GaussFCC2 captures the details of the attack better compared to other features. It also shows that when the device quality is low and the attacker-to-talker distance is varied, GaussFCC1 captures the details of the attack better compared to other features. Figure 4 shows that the analysis energy band for GaussFCC1 is comparatively narrower than the one for GaussFCC2.

5.5 Analysis IV: Performance Analysis of Countermeasure on Female and Male Speaker Utterances

Table 6 is used to show the available male and female utterances in the development and evaluation set of ASVspoof2019 dataset. This information is used for further analysis.

The spoof detection for male and female utterances categorised based on features with both generative and discriminative classifiers is shown in Table 7. In paper [38], for speaker recognition task LFCC is suggested to be the good option especially for female trials. The results as shown in Table 7 in the case of spoof detection task, LFCC and its modified versions GaussFCC1 and GaussFCC2 remain unbiased. The least improvement in female trials could be attributed to the number of female utterances being more than male utterances from Table 6.

5.6 Analysis V: Score-Level Fusion of GaussFCC1 and GaussFCC2

Tables 3, 5 and 7 show that GaussFCC1 and GaussFCC2 features are complementary from GMM classifier under LA condition and from BiLSTM under PA condition. This

Table 7 Performance of countermeasures on female and male utterances under both LA and PA conditions. (**columnwise**) Across classifiers and (**rowwise**) across female and male utterances for both development and evaluation data

Feature (classifier)	Development				Evaluation			
	Female		Male		Female		Male	
	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF
Logical access (LA)								
LFCC(GMM)	4.2881	0.1153	0.0135	0.0003	4.6721	0.1153	1.1646	**0.0294**
GaussFCC1(GMM)	0.0000	**0.0000**	0.0000	**0.0000**	4.6687	0.1145	4.3726	0.1114
GaussFCC2(GMM)	0.000000	**0.0000**	0.0000	**0.0000**	4.8724	0.1164	3.7659	0.0999
LFCC(BiLSTM)	0.0600	0.0016	0.2167	0.0048	7.1790	0.1440	7.0120	0.1466
GaussFCC1(BiLSTM)	0.0600	0.0007	0.2099	0.0041	9.6969	0.1615	8.2678	0.1803
GaussFCC2(BiLSTM)	0.0000	**0.0000**	0.0068	0.0001	7.0777	**0.0986**	5.8677	0.0970
Physical access (PA)								
LFCC(GMM)	10.5556	0.2924	10.7407	0.2254	12.3122	0.2924	12.7893	0.2991
GaussFCC1(GMM)	7.5600	0.1682	8.6548	0.1813	9.5006	0.2222	9.9238	0.2415
GaussFCC2(GMM)	8.9506	0.1948	9.5370	0.2085	10.9916	0.2664	11.6672	0.2875
LFCC(BiLSTM)	1.1437	0.0328	1.7155	0.0435	1.8423	0.0545	2.1480	0.0624
GaussFCC1(BiLSTM)	0.7407	**0.0197**	1.4326	0.0409	1.3693	**0.0413**	1.5307	**0.0457**
GaussFCC2(BiLSTM)	0.7407	0.0207	1.214	**0.0367**	1.4011	0.0425	2.0245	0.0600

Comparatively good performance obtained for the male and female trials for development and evaluation data under LA and PA conditions

Table 8 Score-level fusion of GaussFCC1 and GaussFCC2 from each of GMM and BiLSTM classifier

Classifier (features)	Logical access (LA)				Physical access (PA)			
	Development		Evaluation		Development		Evaluation	
	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF	EER%	min t-DCF
GMM (GaussFCC1)	0.0000	0.0000	4.6088	0.1140	8.1111	0.1795	9.6740	0.2325
GMM (GaussFCC2)	0.0000	0.0000	4.6226	0.1127	9.2788	0.2003	11.2990	0.2768
GMM (GaussFCC1+GaussFCC2)	**0.0000**	**0.000000**	**4.1468**	**0.1016**	8.4424	0.1872	10.1052	0.2446
BiLSTM (GaussFCC1)	0.1127	0.0020	9.2967	0.1691	1.0895	0.0302	1.4761	0.0440
BiLSTM (GaussFCC2)	0.0022	0.000045	6.6653	0.0982	1.0741	0.0304	1.7137	0.0508
BiLSTM (GaussFCC1+GaussFCC2)	0.0308	0.0002	6.8774	0.1121	**0.8693**	**0.0250**	**1.2990**	**0.0385**

The highlighted values show the comparatively good performance achieved by GMM and BiLSTM on the score-level fusion of GaussFCC1 and GaussFCC2 features

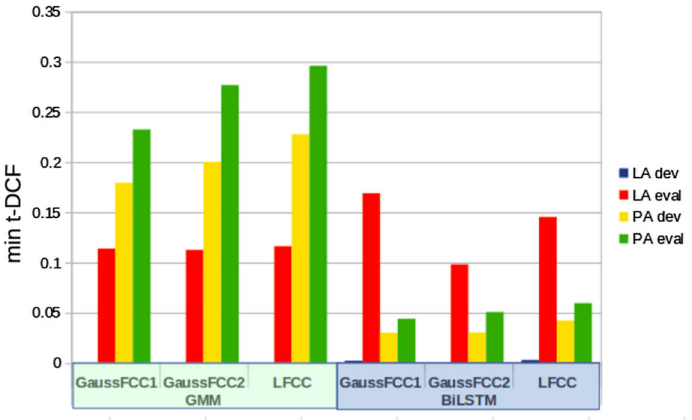

Fig. 9 Performance comparison of baseline and proposed features using generative (GMM) and discriminative (BiLSTM) classifiers

was tested with further analysis using score-level fusion of GaussFCC1 and GaussFCC2 from both the classifiers and hence proves the above observation. The result of the fusion is shown in Table 8. The highlighted values show that the performance is comparatively good. The observations in the paper empirically prove that under LA condition the spoof detection is good using GMM classifier with the proposed feature (in isolation and in combination as well) and under PA condition the spoof detection is good using BiLSTM classifier. In all the experimental outcome, both the proposed features, namely GaussFCC1 and GaussFCC2, have outperformed the baseline feature traditional LFCC.

In a nutshell, the comparative analysis of the countermeasure with GaussFCC features using both GMM and BiLSTM classifiers is depicted in Fig. 9.

The comparison shows the countermeasure performance well in countering attacks under LA scenario when used with GMM classifier and under PA scenario with BiLSTM classifier. To get the best out from these generative and discriminative classifiers as suggested by the authors of [1], the research could be taken forward with classifier-level fusion.

Based on the observations of the empirical outcome, the countermeasure has shown robustness in spoof detection on a speaker verification system using a complementary combination of features and classifiers as discussed. The result shows consistency of performance by the countermeasure at all stages of analysis, namely between features, classifiers and gender-based utterances. Hence leading to the choice of optimal features for score-level fusion and appropriate selection of classifiers for spoof detection in an automatic speaker verification system by experimenting on ASVspoof2019 datasets.

6 Conclusions

The literature bears the witness of good spoof detection in ASV system with the use of linear frequency cepstral coefficients, in isolation and in fusion with other features as

discussed in the paper. The authors intend to explore the possibility to bring enhancement to the linear frequency-based feature so as to use the feature to its full potential in spoof detection. This could be achieved through the modification to the weighting function from triangular to Gaussian. The change in cepstral coefficients obtained showed significant variations. Subsequently the computation of entropy over the probability distribution of Gaussian filtered energy spectrum proved that the changes in cepstral coefficients are due to the information gain achieved using Gaussian filters. Two variations of modified features are GaussFCC1 and GaussFCC2 as discussed. Both these features were used with two different classifiers to detect spoof attack under LA and PA conditions. The two classifiers used are generative (GMM) and discriminative (BiLSTM) classifiers. The paper presents an elaborate analysis of the empirical results with the possible combinations of features and classifiers. A study of these combinations was carried out for female and male speaker's utterances as well. Throughout the analysis, the performance of the countermeasure showed consistent improvement and hence empirically proved that GaussFCC1 and GaussFCC2 are optimal candidates for score-level fusion. The behaviour of generative classifier was found to be good under LA condition and discriminative classifier under PA condition. The utterances considered for the experiments were not pre-emphasised in the time domain as per the reasons discussed in the paper. The pre-processing of speech signals might cause loss/modification of information at each level of processing which could be averted, and from this perspective, the paper investigates the performance of countermeasure to spoof attack with non-pre-emphasised utterance. Further investigation could be carried out by increasing the cepstral coefficients and study the influence of discrete cosine transform on the logarithm of Gaussian filtered energy spectrum and subsequent impact on cepstrum.

Acknowledgements We extend our thanks to SSN College of Engineering for providing us with the required infrastructure to carry out our research work.

References

1. C.M. Bishop, J. Lasserre, Generative or discrimative? Getting the best of both worlds, vol. 8, pp. 3–23 (2007)
2. D.R. Campbell, K.J. Palomäki, G. Brown, A matlab simulation of "shoebox" room acoustics for use in research and teaching. Comput. Inf. Syst. J. **9**(3), 48 (2005). (**ISSN 1352-9404**)
3. K. Conrad, Probability distributions and maximum entropy (2005)
4. R.K. Das, J. Yang, H. Li, Long range acoustic features for spoofed speech detection, in *INTERSPEECH* (2019)
5. L. Deng, D. O'Shaughnessy, Speech processing: a dynamic and optimization-oriented approach. Marcel Dekker Inc., (2003). https://doi.org/10.1201/9781482276237
6. A.R. Douglas, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture models. IEEE Trans. Speech Audio Process. **1**, 72–83 (1995). https://doi.org/10.1109/89.365379
7. S.K. Ergünay, E. Khoury, A. Lazaridis, S. Marcel, On the vulnerability of speaker verification to realistic voice spoofing (2015), pp. 1–6. https://doi.org/10.1109/BTAS.2015.7358783
8. M.D. Femila, A.A. Irudhayaraj, Biometric system. in *2011 3rd International Conference on Electronics Computer Technology*, , vol 1, pp. 152–156 (2011). https://doi.org/10.1109/ICECTECH.2011.5941580
9. C. Hanilci, T. Kinnunen, Md. Sahidullah, A. Sizov, Classifiers for synthetic speech detection: a comparison, in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)* (2015), pp. 2057–2061

10. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735 (**ISSN 0899-7667**)
11. M.R. Kamble, H.A. Patil, Analysis of reverberation via Teager energy features for replay spoof speech detection, in *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 2607–2611. https://doi.org/10.1109/ICASSP.2019.8683830
12. T. Kinnunen, Z. Wu, K.A. Lee, F. Sedlak, E.S. Chng, H. Li, Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 4401—4404
13. T. Kinnunen, K.A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, D.A. Reynolds, t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, in *Proceedings*, Odyssey 2018 (2018)
14. M. Kudo, J. Toyama, M. Shimbo, Multidimensional curve classification using passing-through regions. Pattern Recognit. Lett. **20**(11), 1103–1111 (1999). https://doi.org/10.1016/S0167-8655(99)00077-X (**ISSN 0167-8655**)
15. M.G. Kumar, Suvidha Rupesh Kumar, M.S. Saranya, B. Bharathi, H.A. Murthy, Spoof detection using time-delay shallow neural network and feature switching, in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019). https://doi.org/10.1109/asru46091.2019.9003824
16. Suvidha Rupesh Kumar, B. Bharathi, A novel approach towards generalization of countermeasure for spoofing attack on ASV systems. Circuits Syst. Signal Process. **40**, 872–889 (2021). https://doi.org/10.1007/s00034-020-01501-y (**ISSN 1531-5878**)
17. O. Kwon, I. Jang, C. Ahn, H. Kang. Emotional speech synthesis based on style embedded tacotron2 framework (2019), pp. 1–4. https://doi.org/10.1109/ITC-CSCC.2019.8793393
18. X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, H. Meng, Replay and synthetic speech detection with res2net architecture In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6–11, 2021, pp. 6354–6358. IEEE (2021). https://doi.org/10.1109/ICASSP39728.2021.9413828
19. D. Mitrovic, M. Zeppelzauer, C. Breiteneder, Features for content-based audio retrieval. Adv. Comput. **78**, 71–150 (2010). https://doi.org/10.1016/S0065-2458(10)78003-7
20. A. Novak, P. Lotton, L. Simon, Synchronized swept-sine: theory, application, and implementation. J. Audio Eng. Soc. **63**(10), 786–798 (2015). (**ISSN 1352-9404**)
21. S.P. Panda, Intelligent voice-based authentication system (2019), pp. 757–760. https://doi.org/10.1109/I-SMAC47947.2019.9032671
22. Y. Qian, N. Chen, K. Yu, Deep features for automatic spoofing detection. Speech Commun. **85**, 43–52 (2016). https://doi.org/10.1016/j.specom.2016.10.007
23. R.A. Rashid, N.H. Mahalin, M.A. Sarijari, A.A. Abdul Aziz, Security system using biometric technology: design and implementation of voice recognition system (VRS) (2008), pp. 898–902. https://doi.org/10.1109/ICCCE.2008.4580735
24. Md. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in *Interspeech* (2015), pp. 2087–2091
25. T.J. Sefara, T.B. Mokgonyane, M.J. Manamela, T.I. Modipa, Hmm-based speech synthesis system incorporated with language identification for low-resourced languages (2019), pp. 1–6. https://doi.org/10.1109/ICABCD.2019.8851055
26. C.E. Shannon, W. Weaver, *A Mathematical Theory of Communication* (University of Illinois Press, Illinois, 1963). (**ISBN 0252725484**)
27. K. Sriskandaraja, V. Sethu, P.N. Le, E. Ambikairajah, Investigation of sub-band discriminative information between spoofed and genuine speech. Interspeech **2016**, 1710–1714 (2016)
28. M. Todisco, H. Delgado, N. Evans, A new feature for automatic speaker verification anti-spoofing: constant q cepstral coefficients, in *Proceedings of the Speaker and Language Recognition Workshop* (2016), pp. 283–290
29. M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K.A. Lee, Asvspoof 2019: future horizons in spoofed and fake audio detection, in *Interspeech 2019* (2019)
30. E. Vincent. Roomsimove (2008). http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip
31. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: a survey. Speech Commun. **66**, 130–153 (2015). https://doi.org/10.1016/j.specom.2014.10.005 (**ISSN 0167-6393**)

32. Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, M. Hanil çi, C. Sahidullah, A. Sizov, Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in *Interspeech* (2015), pp. 2037–2041
33. Z. Xie, W. Zhang, Z. Chen, X. Xu, A comparison of features for replay attack detection. J. Phys. Conf. Ser. (JPCS) **1229**, 8 (2019)
34. J. Yamagishi, M. Todisco, Md. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K.A. Lee, V. Vestman, A. Nautsch, Asvspoof 2019: automatic speaker verification spoofing and countermeasures challenge evaluation plan (2019)
35. J. Yang, L. Xu, B. Ren, Y. Ji, Discriminative features based on modified log magnitude spectrum for playback speech detection. EURASIP J. Audio Speech Music Process. (2020). https://doi.org/10.1186/s13636-020-00173-5 (**ISSN 1352-9404**)
36. H. Yu, Z.H. Tan, Y. Zhang, Z. Ma, J. Guo, Dnn filter bank cepstral coefficients for spoofing detection. IEEE Access **5**, 4779–4787 (2017)
37. C. Zhang, C. Yu, J.H.L. Hansen, An investigation of deep-learning frameworks for speaker verification antispoofing. IEEE J. Sel. Top. Signal Process. **11**(4), 684–694 (2017). https://doi.org/10.1109/JSTSP.2016.2647199
38. X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, Linear versus mel frequency cepstral coefficients for speaker recognition, in 2011 *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 559–564 (2011). https://doi.org/10.1109/ASRU.2011.6163888