Speech Frame Selection for Spoofing Detection with an Application to Partially Spoofed Audio-Data

Kishore A. Kumar, Dipjyoti Paul, Monisankha Pal, Md Sahidullah, Goutam Saha

Noname manuscript No.
(will be inserted by the editor)

Speech Frame Selection for Spoofing Detection with an Application to Partially Spoofed Audio-Data

A Kishore Kumar [1] · Dipjyoti Paul [2] · Monisankha Pal [3] · Md Sahidullah [4] · Goutam Saha [1]

the date of receipt and acceptance should be inserted later

Abstract In this paper, we introduce a frame selection strategy for improved detection of spoofed speech. A countermeasure (CM) system typically uses a Gaussian mixture model (GMM) based classifier for computing the log-likelihood scores. The average log-likelihood ratio for all speech frames of a test utterance is calculated as the score for the decision making. As opposed to this standard approach, we propose to use selected speech frames of the test utterance for scoring. We present two simple and computationally efficient frame selection strategies based on the log-likelihood ratios of the individual frames. The performance is evaluated with constant-Q cepstral coefficients as front-end feature extraction and two-class GMM as a back-end classifier. We conduct the experiments using the speech corpora from ASVspoof 2015, 2017, and 2019 challenges. The experimental results show that the proposed scoring techniques substantially outperform the conventional scoring technique for both the development and evaluation data set of ASVspoof 2015 corpus. We did not observe noticeable performance gain in ASVspoof 2017 and ASVspoof 2019 corpus. We further conducted experiments with partially spoofed data where spoofed data is created by augmenting natural and spoofed speech. In this scenario, the proposed methods demonstrate considerable performance improvement over baseline.

[1] Department of Electronics & ECE, Indian Institute of Technology Kharagpur, India.
E-mail: kishore@iitkgp.ac.in, gsaha@ece.iitkgp.ac.in
[2] Department of Computer Science, University of Crete, Greece.
E-mail: dipjyotipaul@csd.uoc.gr
[3] Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, USA.
E-mail: mp_323@usc.edu
[4] Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France.
E-mail: md.sahidullah@inria.fr

Keywords Anti-spoofing · ASVspoof · Countermeasures · Frame selection · Partially spoofed speech · Speaker verification · Synthetic speech detection · Voice conversion

1 Introduction

The voice-based authentication using automatic speaker verification (ASV) technology is highly vulnerable to the *spoofing attacks* with speech signals generated using *voice conversion* (VC), *speech synthesis* (SS), and *replay* method (Wu et al. 2015). Detection of spoofed voices is the most important concern for the development of spoofing countermeasures. Over the last few years, significant efforts have been devoted to designing different countermeasures (CM) to improve the security of voice biometric systems. Different features and classifiers are investigated for this task (Sahidullah et al. 2019). Along with the popular speech features used in other speech applications, studies have been conducted for better representation of the speech signal for the spoofed speech detection task (Sahidullah et al. 2015; Paul et al. 2017; Todisco et al. 2017; Patel and Patil 2017; Pal et al. 2018). Similarly, different back-end classifiers are also designed to improve spoofing detection performance (Hanilçi et al. 2015; Villalba et al. 2015b; Tian et al. 2016). The works in (Sahidullah et al. 2019; Kamble et al. 2020) reported up-to-date reviews of recently developed spoofing countermeasure methods.

Recently several deep neural network (DNN) based CMs are proposed and reported significant performance improvement (Villalba et al. 2015b; Tian et al. 2016; Yu et al. 2017). However, simple modeling techniques such as the Gaussian mixture model (GMM), the trained models store a fewer model parameters, as opposed to the case of DNN-based approaches, where the number of model parameters went up to some hundreds of thousands which consumes considerable size of physical memory. The simple modelling techniques are favourable for practical purpose specially when the computational resources are limited and the storage requirement is an issue. This work improves the standard GMM-based spoofing countermeasures with an improved scoring technique. The existing works on spoofing detection use all the speech frames from a speech signal to model the natural and synthetic speech class. While testing, all the speech frames of the test utterance are used in computing the detection score. Considering all the available speech frames with equal weight is also a common practice in ASV (Reynolds and Rose 1995). Recently, *attention modeling* has shown promising improvement in ASV performance where the contributions from the speech frames are weighted and combined according to their importance (Zhu et al. 2018; Okabe et al. 2018). A DNN trained with attention mechanism has helped to accurately detect replay-based spoofed voice in version 1.0 of the ASVspoof 2017 dataset (Tom et al. 2018)[1]. Based on these studies, we hypothesize that utilizing all the speech frames may

[1] The natural speech files in this version of the dataset contain some zero-sequence artifacts at the beginning which might help in the detection process with attention model.

not be a good choice for spoofing detection task. In the voice-spoofing process also, all the speech frames are not necessarily spoofed. Usually, during voice conversion, the voiced frames are only transformed whereas unvoiced frames are copied from source speech frames (Erro et al. 2010). As a result, the training and scoring processes are affected by the proportion of the converted and unconverted frames in the spoofed speech utterance. Thereby, an utterance with a smaller fraction of converted frames is more likely to be detected as a genuine speech by the spoofing detector. A similar problem can also arise for spoofed speech signal generated using the speech synthesis method. For example, in *unit selection* based approach, the frames in the unit boundaries have relatively more artifacts whereas the individual units are very similar to a natural voice (Tian et al. 2016). This also makes the spoofing detection task more challenging– for example detection of MaryTTS-based attack in the ASVspoof 2015 corpus (Wu et al. 2017). Our work investigates the use of selected frames in spoofing detection task.

The use of selected frames has been found useful for several speech processing applications. The most common practice is the usage of *speech activity detector* (SAD) to discard unreliable speech frames in speech and speaker recognition task. Speech frames are also selectively used to speed up the computational time in real-time speaker recognition (Kinnunen et al. 2006). In (Kwon and Narayanan 2007), discriminative speech frames identified by the likelihood ratio based approach are utilized for speaker identification task. In another work (Jung et al. 2010), mutual information based frame selection is proposed for speaker recognition task, where speech frames with minimum-relevancy within selected feature frames but maximum-relevancy to speaker models are used. The authors in (Fujihara et al. 2010) utilized reliable frames for modeling the characteristics of singing voice where the unreliable speech region consisting of non-vocal sounds are discarded. The non-speech frames are found useful in cell-phone recognition where frames are identified using an energy-based SAD (Hanilçi and Kinnunen 2014). In a recent work (Ventura et al. 2015), speech frames with higher magnitude are used for bird sound identification. To exploit the effects of long and short-duration artifacts, weighted likelihood-ratio score based approach is also proposed for spoofing detection in (Khodabakhsh and Demiroglu 2016).

Other than the use of SAD for discarding non-speech frames (Villalba et al. 2015a; Jahangir et al. 2015), our work is the first attempt to explore the *frame selection* method for voice spoofing detection task. The previous study reveals that non-speech frames can also be useful for synthetic speech detection (Sahidullah et al. 2015). Therefore, we do not reject explicitly the non-speech frames. Rather we first study different methods for finding potentially relevant speech frames from all the speech frames. Since the indication about spoofing from fewer frames could be sufficient for CM task, the motivation of this work is to use more informative and reliable speech frames in the final scoring. The experiments are conducted on three ASVspoof databases (Wu et al. 2017; Kinnunen et al. 2017a). Moreover, the idea is evaluated against a realistic condition where the data is partially spoofed when natural speech is

augmented with synthetic or replay speech. We consider a possible scenario where the intruder has access to a small segment of a digital copy of the target speaker's speech. The intruder further concatenates it with spoofed speech and try to access the system protected with voice biometrics[2].

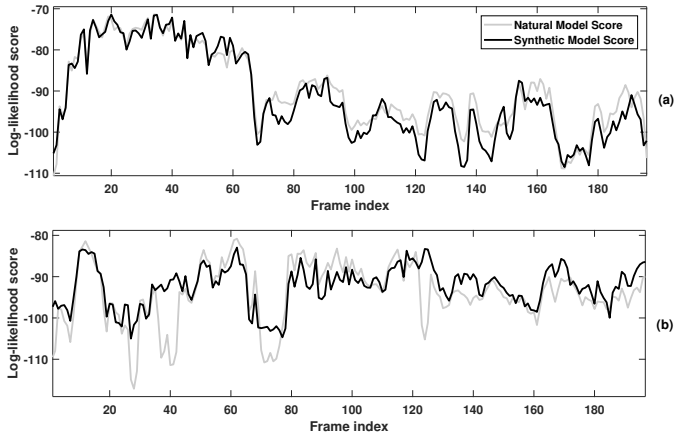

Fig. 1 Frame-level log-likelihood scores of (a) natural speech signal, and (b) synthetic speech signal computed against natural and synthetic model.

2 Speech Frame Selection for Anti-spoofing

2.1 Motivation and Background

A spoofing countermeasure system discriminates between natural and synthetic speech signal. Conventionally, the system computes average log-likelihood score over all the speech frames of an utterance for given models, *i.e.*, natural and synthetic. The higher the average score against a particular model, it is more likely to classify the given sample belonging to that class. Though the speech frames created from an utterance collectively show a trend in overall log-likelihood ratio, decisions obtained from individual speech frames can vary from frame-to-frame. For example, some frames in a speech utterance may have a higher log-likelihood score against natural class and some frames may have higher scores against synthetic class. Fig. 1(a) and 1(b) illustrate the frame-level scores of a natural and a synthetic speech signal respectively, against both natural and synthetic models. We observe that the likelihood of natural speech signal frames for the natural model is higher than the synthetic

[2] Other than voice-biometrics, this situation may also encounter where someone creates fake speech by combining segments from multiple sources.

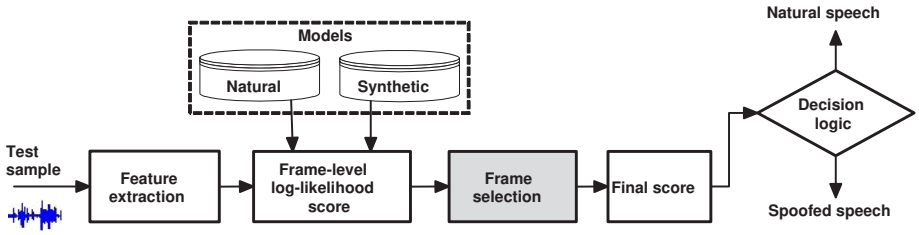

Fig. 2 Block diagram of the proposed evaluation framework.

model in most cases but not always. Similarly, for scores of synthetic speech frames in Fig. 1(b), we notice that some frames show a higher likelihood for natural class, too. Therefore, a suitable frame selection criteria that helps to select important speech frames could be useful. We propose two schemes in the next subsection.

2.2 Proposed Frame Selection Technique

In a GMM-based spoofing detector, the log-likelihood score is calculated as $\Lambda(\mathbf{X}) = \mathcal{L}(\mathbf{X}|\lambda_n) - \mathcal{L}(\mathbf{X}|\lambda_s)$. Here, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is the feature matrix of the test utterance where T is the number of frames and $\mathcal{L}(\mathbf{X}|\lambda)$ is the average log-likelihood of $\mathbf{X}$ given GMM model λ. λ_n and λ_s are the natural and synthetic models, respectively. $\mathcal{L}(\mathbf{X}|\lambda)$ is defined by the following equation,

$$\mathcal{L}(\mathbf{X}|\lambda) = \frac{1}{T}\sum_{i=1}^{T} \log p(\mathbf{x}_i|\lambda). \tag{1}$$

We introduce a frame selection scheme which only uses selected speech frames in the final score computation, illustrated in Fig. 2. We propose to select the frames, based on their likelihood ratio, i.e., a frame i is selected based on $p(\mathbf{x}_i|\lambda_n)$ and $p(\mathbf{x}_i|\lambda_s)$. The steps for computing relevant frames are as follows:

Step I: compute $\log p(\mathbf{x}_i|\lambda_n)$ and $\log p(\mathbf{x}_i|\lambda_s)$ for all T frames in an utterance.

Step II: compute the likelihood ratio l_i for all T frames as,

$$l_i = \log p(\mathbf{x}_i|\lambda_n) - \log p(\mathbf{x}_i|\lambda_s). \tag{2}$$

Step III: set a threshold θ. A frame, i, is retained for scoring if $l_i < \theta$, otherwise, it is discarded.

Threshold Selection: The selection of threshold, θ, is an important task in the above stated frame-selection process. We propose two different methods for this purpose. The first method is to set θ as zero to select frames with higher likelihood for spoofing class for final score computation. The second method uses an utterance-dependent threshold selection scheme, where the threshold

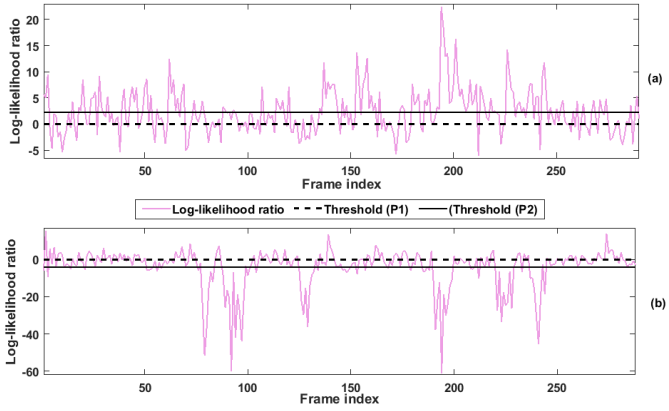

Fig. 3 Selection of speech frames for proposed two methods based on log-likelihood ratio scores of (a) natural and (b) spoofed speech signal. The horizontal lines correspond to threshold lines for $P1$ (dotted) and $P2$ (continuous) method. Speech frames with log-likelihood ratio lower than the threshold line are retained for final scoring.

is set as average of l_i computed over all the frames, i.e., $\theta = \frac{1}{T}\sum_i l_i$. In this work, we call the zero threshold-based approach as P1 and the mean-based approaches as P2.

Fig. 3 shows the illustration of selected frames from both proposed methods for a single speech utterance of natural (Fig. 3(a)) and synthetic speech (Fig. 3(b)) when during test. This indicates that if the test speech is natural, P2 method considers a larger number of frames than P1. On the other hand, if the test speech is synthetic, a smaller number of frames is retained by P2 method corresponding to a higher similarity with synthetic speech class.

3 Experimental Setup

3.1 Database Description

3.1.1 ASVspoof Corpora

We evaluate the spoofing detection performance on three different corpora: ASVspoof 2015 (Wu et al. 2015), ASVspoof 2017 (Kinnunen et al. 2017b) and ASVspoof 2019 (Todisco et al. 2019). The first database consists of seven different VC (S1, S2, S5, S6, S7, S8, and S9) and three SS-based (S3, S4, and S10) spoofing techniques while the latter consists of replay attacks collected from the *wild conditions*. ASVspoof 2019 corpus is divided into two data sets: logical access (LA) and physical access (PA). LA data condition consists of spoofed data from TTS and VC based techniques, whereas, in PA, spoofed data from replay attack is considered. The challenge database is based upon a

standard multi-speaker speech synthesis database called VCTK[3]. The spoofed utterances were prepared using 19 different TTS and VC based techniques for LA and nine replay configurations for PA. All the databases have three subsets: *training*, *development* and *evaluation*. The spoofing countermeasure systems are trained on the training subset and evaluated on the other two subsets. The number of genuine and spoofed utterances in the training, development and evaluation subset of all three databases are summarized in Table 1 and 2.

Table 1 Number of utterances in training, development and evaluation set of ASVspoof 2015 and 2017 corpus.

	ASVspoof 2015		**ASVspoof 2017**	
	Natural	**Spoofed**	**Natural**	**Spoofed**
Training	3750	12625	1507	1507
Development	3497	49875	760	950
Evaluation	9404	184000	1298	12008
Total	**16651**	**246500**	**3565**	**14465**

Table 2 Number of utterances in training, development and evaluation subsets of ASVspoof 2019 database

Subset	**#utterances**			
	LA		**PA**	
	Natural	**Spoof**	**Natural**	**Spoof**
Training	2580	22800	5400	48600
Development	2548	22296	5400	24300
Evaluation	7355	63882	18090	116640

3.1.2 Partially Spoofed Data

We have designed conditions where partially spoofed data is used for testing. To simulate this test condition, we augment spoofed data with speech segment consisting of natural voice. For experiments with synthetic speech, we augment synthetic speech in different proportion with natural speech files in the development set of ASVspoof 2015. Similarly, for experiments with replayed speech in ASVspoof 2017, we extend each natural speech file by concatenating spoofed speech. We randomly pick a spoofed speech file from the same speaker's data in the corresponding dataset. The steps involved in producing the partially spoofed data are presented in Algorithm 1 and Fig. 4. A comparison of the speech spectrum of a natural and 40% spoofed speech signal is presented in Fig. 5. In contrast to the original protocol, we have kept same number of speech

[3] http://dx.doi.org/10.7488/ds/1994

files for natural and spoofed data in the experiments with partially spoofed data.

Algorithm 1: Preparation of partially spoofed data

function DataPreparation()
N = No. of natural files
while $i <= N$ **do**
  S_{nat} = read(Natural Speech File)
  L_{nat} = length(S_{nat})
  find(Spoof file, where $L_{\text{spoof}} >= \alpha * L_{\text{nat}}$, from the same speaker)
  if *file found* **then**
    S_{spoof} = read(Spoofed Speech File)
    $S_{\text{Spoof},\alpha}$ = concatenate($S_{\text{nat}}, S_{\text{spoof}}[1 : \alpha * L_{\text{nat}}]$)
  else
    break

Where L_{nat} and L_{spoof} are length of natural and spoof speech files respectively, and α is the factor by which data is partially spoofed.

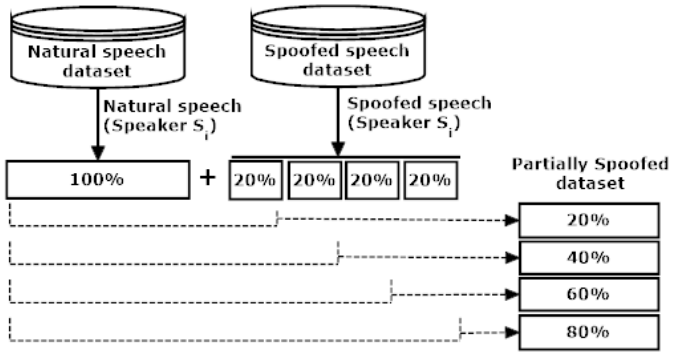

Fig. 4 Illustration showing the data preparation of partially spoofed speech dataset

3.2 Feature Extraction and Classifier

We evaluate the proposed methods with two different acoustic features. The first one is *mel-frequency cepstral coefficients* (MFCCs), the most widely used features for speech processing applications. The MFCCs are extracted using 20 filters in mel scale. We augment the first and second order dynamic coefficients (i.e., delta and double-delta) to form 60-dimensional cepstral features (Paul et al. 2017).

The other feature we used is CQCC. This feature was used with GMM back-end to produce state-of-the-art spoofing detection performance in the

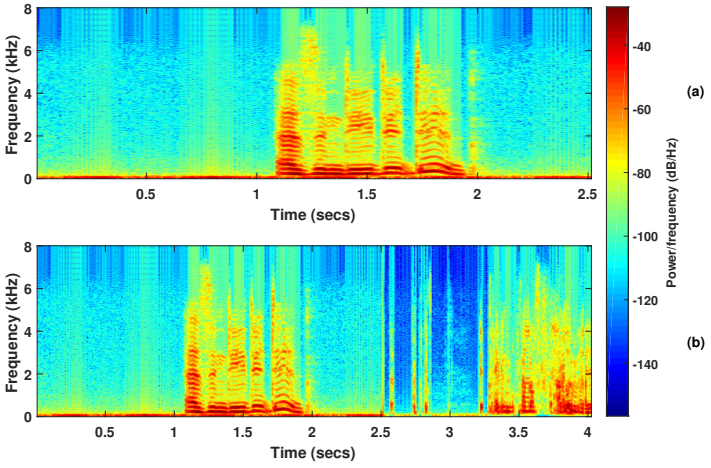

Fig. 5 Comparison of spectrum of (a) natural and (b) 40% partially-spoofed natural speech signal.

chosen datasets (Delgado et al. 2018; Todisco et al. 2017). Unlike short-time Fourier transform which provides fixed time-frequency resolution and is used in MFCC formulation, the constant-Q transform (CQT) used in CQCC extraction process provides a higher frequency resolution for the lower frequencies and a higher temporal resolution for the higher frequencies. The CQT performs perceptually motivated wavelet-like time-frequency analysis that uses a constant-Q factor across the entire spectrum by employing geometrically spaced frequency bins. While calculating cepstral features, a spline interpolation method is applied to *resample* the geometric frequency scale into a uniform linear scale for applying linearly-spaced DCT coefficients (Todisco et al. 2017). We use 60-dimensional CQCC features consisting of 20-dimensional static coefficients augmented with dynamic coefficients.

We train the back-end GMM with 512 mixture components by *maximum-likelihood* criterion using ten iterations of *expectation-maximization* (EM) algorithms. We use similar acoustic features and classifiers for both synthetic speech detection and replay attack detection task.

3.3 Performance Evaluation

We use *equal error rate* (EER) as the evaluation metric to assess the spoofing countermeasures performance. We calculate EER using BOSARIS toolkit [4] which uses *receiver operating characteristics convex hull* (ROCCH) method. A lower value of EER indicates a better performance.

[4] https://sites.google.com/site/bosaristoolkit/

Table 3 Performance (in % of EER) comparison of Baseline ($B1$) and Baseline-SAD ($B2$) with proposed $P1$ and $P2$ selection criteria for MFCC and CQCC feature on ASVspoof 2015 development data.

		S1	**S2**	**S3**	**S4**	**S5**	**Avg.**
MFCC	**B1**	0.1336	3.2519	**0.0000**	**0.0000**	2.174	1.1119
	B2	**0.0354**	3.0930	0.0614	0.0413	0.8982	0.8258
	P1	0.1461	**1.1036**	**0.0000**	**0.0000**	**0.8105**	0.4120
	P2	0.0417	1.1844	**0.0000**	**0.0000**	0.8269	**0.4106**
CQCC	**B1**	**0.0000**	0.2534	**0.0000**	**0.0000**	0.8056	0.2118
	B2	**0.0000**	**0.0284**	0.0897	0.0074	0.2648	0.0780
	P1	0.0992	0.4162	0.0614	0.0779	0.3708	0.2051
	P2	**0.0000**	0.0825	**0.0000**	**0.0000**	**0.1835**	**0.0532**

Table 4 Performance (in % of EER) comparison of Baseline ($B1$) and Baseline-SAD ($B2$) with proposed $P1$ and $P2$ selection criteria for MFCC and CQCC feature on ASVspoof 2015 evaluation data.

		MFCC				**CQCC**			
		B1	**B2**	**P1**	**P2**	**B1**	**B2**	**P1**	**P2**
Known Attack	**S1**	0.0360	0.0312	0.1116	**0.0083**	0.0072	**0.0064**	0.0711	0.0104
	S2	2.5410	3.0854	1.1816	**0.7223**	0.2318	**0.0198**	0.4024	0.0756
	S3	**0.0000**	0.0309	**0.0000**	**0.0000**	**0.0000**	0.0348	0.0209	**0.0000**
	S4	**0.0000**	0.0295	**0.0000**	**0.0000**	**0.0000**	0.0179	0.0210	**0.0000**
	S5	1.5548	0.9351	0.9372	**0.4275**	0.5451	**0.0992**	0.3663	0.1726
	Avg.	0.8263	0.8224	0.4460	**0.2316**	0.1568	**0.0357**	0.1763	0.0517
Unknown Attack	**S6**	1.5250	1.2144	0.7821	**0.3253**	0.3633	**0.0615**	0.3125	0.0700
	S7	0.2770	0.1295	0.0910	**0.0253**	0.0452	**0.0000**	0.1355	0.0169
	S8	0.0518	0.3366	0.2362	**0.0064**	**0.0152**	0.0532	1.0640	0.0365
	S9	0.2965	0.0408	0.1516	**0.0180**	0.0688	**0.0000**	0.2200	0.0193
	S10	23.6808	30.6837	**6.0007**	11.3609	4.4230	10.2633	**0.2759**	0.7490
	Avg.	5.1662	6.4810	**1.4523**	2.3471	0.9831	2.0756	0.4016	**0.1784**
Avg.		2.9963	3.6517	**0.9492**	1.2894	0.5699	1.0600	0.2889	**0.1150**

4 Results & Discussion

4.1 Experiments on ASVspoof 2015

First, we evaluate the spoofing detection performance on ASVspoof 2015 corpus. Table 3 and 4 show the results on development and evaluation set, respectively. In our case baseline system is the one which considers all the speech frames during score calculation. We have compared the performance of the proposed methods with the baseline system ($B1$) as well as the baseline with SAD ($B2$) based frame selection approach, where the non-speech frames are discarded only during the test. We use an energy-based SAD where the first coefficient of CQCC feature after DCT is used as energy, and the decision threshold is obtained by performing k-means clustering on the log-energies

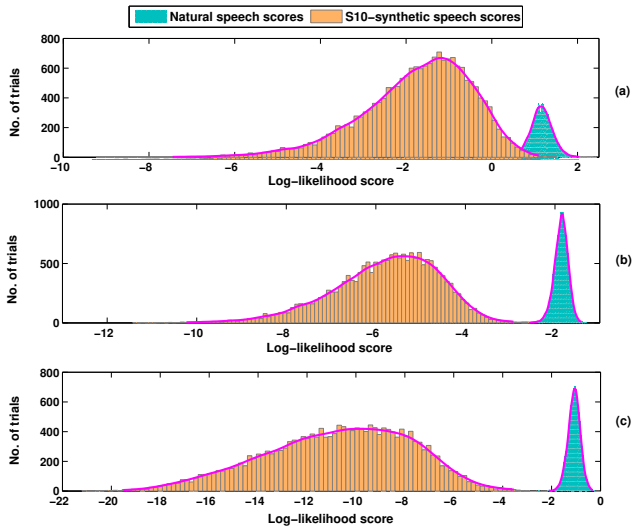

Fig. 6 Distribution of spoofing detection scores for natural and synthetic (only S10) speech files from ASVspoof 2015 database for (a) baseline, (b) P1, and (c) P2 methods for CQCC feature.

followed by computation of the threshold as mean of the clusters. The experimental results indicate that the average EER for the proposed P2 method that uses the threshold as mean outperforms other techniques for both development and evaluation set. Especially for S5 (a voice conversion technique based on Festvox (Wu et al. 2015)), the relative improvement of P2 method over baseline without SAD is considerably high. This is probably due to the presence of a large number of unvoiced and non-speech frames in spoofed data which are not processed by the S5 method and are discarded with the help of frame selection.

Fig. 3 illustrates the frame selection threshold for a natural and a synthetic speech signal generated with S5 method. The plot shows the log-likelihood ratio for each speech signal and the horizontal lines correspond to threshold values for P1 (dotted) and P2 (continuous) method. From Fig. 3(a), we deduce that the P2 method selects a higher number of frames for natural speech signal than that of synthetic speech signal and it leads to increase in the average log-likelihood ratio of selected frames. On the other hand, for a synthetic speech file, P2 method selects smaller number of frames showing a lower final average likelihood ratio. Both of these help in substantially improving the spoofing detection performance by selecting more informative speech frames. In Fig. 6, we have shown the score distributions of natural and S10 synthetic speech computed for all files from evaluation set of ASVspoof 2015 database. The figure indicates, unlike baseline method, proposed P1 and P2 methods have clear separation of scores between natural and synthetic files.

The baseline system with SAD performs well compared to the method without SAD. However, it exhibits very poor performance for S10. On the other hand, both the proposed methods perform better than baselines when average

EER is computed over all the 10 attack conditions. Interestingly, improvement is noticeably higher for the most difficult attack S10. In this case, the baseline system gives average EER of 4.4230% whereas the P1 and P2 show 0.2759% and 0.7490%, respectively. This is expected as S10 is a unit selection based speech synthesis method which concatenates *diphones* of natural speech to obtain the synthetic speech and spoofed speech related information is retained mainly in the frames in the concatenation points.

4.2 Experiments on ASVspoof 2017

We have shown the spoofing detection results on ASVspoof 2017 in Table 5 on CQCC features. The results indicate that for replay attack, frame selection methods including SAD do not improve the EER over baseline. Moreover, performance is slightly degraded in most cases. This is expected as all the speech frames in spoofed data are converted in playback voice. The frame selection methods, which select speech frames with a higher likelihood of being spoofed, are not applicable for tackling replay attack.

Table 5 Performance (in % of EER) comparison of Baseline ($B1$) and Baseline-SAD ($B2$) with proposed $P1$ and $P2$ selection criteria on ASVspoof 2017 development and evaluation data with CQCC feature.

	Development	**Evaluation**
B1	14.3711	**29.7168**
B2	14.9000	31.3153
P1	16.6962	31.5703
P2	**14.2912**	30.7852

4.3 Experiments on ASVspoof 2019

ASVspoof 2019 corpus was made up of spoofing data from advanced TTS or VC systems and replay speech built in controlled environment, so that they are indistinguishable perceptually from bona fide speech. The performance of our proposed scheme is evaluated against this database and the results are tabulated in Table 6. It is observed that except PA data condition in evaluation set for all other data conditions the proposed system underperforms the baseline system.

4.4 Experiments with Partially Spoofed Test Data

The experimental results for partially spoofed test data are shown in Table 7 for synthetic speech (ASVspoof2015) and replay speech (ASVspoof2017).

Table 6 Performance (in % of EER) comparison of Baseline ($B1$) with proposed $P1$ and $P2$ selection criteria on ASVspoof 2019 development and evaluation data with CQCC feature.

	Development		**Evaluation**	
	Logical Access	**Physical Access**	**Logical Access**	**Physical Access**
B1	**0.4300**	**9.8700**	**9.5700**	11.0400
P1	3.0980	10.5556	21.3059	12.3315
P2	0.5506	9.7778	10.5507	**10.9508**

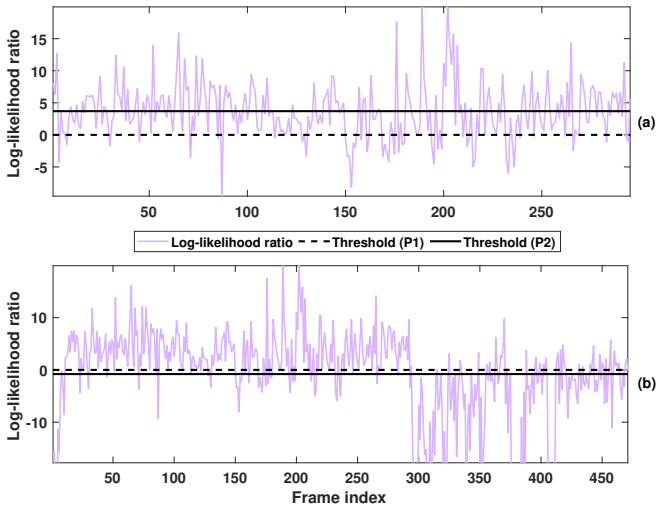

Fig. 7 Selection of speech frames for proposed two methods based on log-likelihood ratio scores of (a) natural and (b) 40% partially spoofed speech signal. The horizontal lines correspond to threshold for $P1$ (dotted) and $P2$ (continuous) method. Speech frames with log-likelihood ratio lower than the threshold are retained for final scoring.

For this experiments, we have used CQCC features as acoustic front-end and GMM-based back-end. We use the models trained on the standard spoofed data. The Table 7 shows that both the proposed methods achieve a considerable performance gain for both type of partially spoofed audio-data. The relative performance gain with proposed approaches over baseline B1 method is more when the amount of spoofed data is higher. For example, with 20% spoofed speech, B1 method shows EER of 28.2109% while the proposed two methods P1 and P2 achieve EERs of 14.9096% and 22.6613%, respectively. On the other hand, for 80% spoofed speech, B1 method gives EER of 5.4413% whereas P1 and P2 method shows EERs of 0.9651% and 1.8797%, respectively. We also observe that the performance gain for partially spoofed created with ASVspoof 2015 is more compared to the partially spoofed data created with ASVspoof 2017 corpus. We further notice that proposed method P1 consistently outperformed the other proposed method P2 for both ASVspoof 2015 and ASVspoof 2017. In Fig. 7, we have compared the selection of frames for

proposed two methods for natural and 40% partial spoofed speech signal. We observe that with higher threshold P1 method selects larger number of spoofed frames than P2 method (Fig. 7(b)).

Table 7 Performance (in % of EER) of Baseline ($B1$) and proposed $P1$ and $P2$ selection criteria on partially spoofed test data created from ASVspoof 2015 and ASVspoof 2017 development set using CQCC features.

		Percentage of augmented spoofed speech			
		20%	**40%**	**60%**	**80%**
ASVspoof 2015	**B1**	28.2109	18.4683	10.2421	5.4413
	P1	**14.9096**	**7.9258**	**2.7667**	**0.9651**
	P2	22.6613	13.0697	5.1207	1.8797
ASVspoof 2017	**B1**	40.5725	34.4875	30.4772	27.5314
	P1	**31.4990**	**24.2021**	**22.6444**	**21.6006**
	P2	37.6820	30.0648	25.3742	22.6008

5 Acknowledgements

We would like acknowledge the funding agencies. Dipjyoti Paul's work is funded by the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu). The work of Md Sahidullah is supported by Region Grand Est.

6 Conclusion

In this paper, we introduced a frame selection strategy for efficient scoring in spoofing detection. The proposed technique is simple and straightforward and does not require much additional computational overhead over existing method using CQCC feature and GMM back-end. The proposed algorithm is a modification of conventional score calculation during testing phase. It shows promising accuracy for synthetic speech detection task on both VC and SS-based spoofed data in ASVspoof 2015. We further evaluated the proposed method for a possible scenario when the test speech is partially spoofed. We observed considerable performance gain for such test conditions. In the current work, we have not modified the training process. This work can be extended by modifying the model training with the help of frame-selection. One can also explore the integration of the frame-selection strategies with latest DNN-based systems.

References

H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, J. Yamagishi, ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements, in *Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018

D. Erro, A. Moreno, A. Bonafonte, Voice conversion based on weighted frequency warping. IEEE Transactions on Audio, Speech, and Language Processing **18**(5), 922–931 (2010)

H. Fujihara, M. Goto, T. Kitahara, H.G. Okuno, A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. IEEE Transactions on Audio, Speech, and Language Processing **18**(3), 638–648 (2010)

C. Hanilçi, T. Kinnunen, Source cell-phone recognition from recorded speech using non-speech segments. Digital Signal Processing **35**, 75–85 (2014)

C. Hanilçi, T. Kinnunen, M. Sahidullah, A. Sizov, Classifiers for synthetic speech detection: A comparison, in *Proc. INTERSPEECH*, 2015, pp. 2057–2061

M.J. Jahangir, P. Kenny, G. Bhattacharya, T. Stafylakis, Development of CRIM System for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015, in *Proc. INTERSPEECH*, 2015, pp. 2072–2076

C.S. Jung, M.Y. Kim, H.G. Kang, Selecting feature frames for automatic speaker recognition using mutual information. IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1332–1340 (2010)

M.R. Kamble, H.B. Sailor, H.A. Patil, H. Li, Advances in anti-spoofing: from the perspective of ASVspoof challenges. APSIPA Transactions on Signal and Information Processing **9** (2020)

A. Khodabakhsh, C. Demiroglu, Investigation of synthetic speech detection using frame-and segment-specific importance weighting. arXiv preprint arXiv:1610.03009 (2016)

T. Kinnunen, E. Karpov, P. Franti, Real-time speaker identification and verification. IEEE Transactions on Audio, Speech, and Language Processing **14**(1), 277–288 (2006)

T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K.A. Lee, The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection, in *Proc. INTERSPEECH*, 2017a, pp. 2–6

T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K.A. Lee, The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017b)

S. Kwon, S. Narayanan, Robust speaker identification based on selective use of feature vectors. Pattern Recognition Letters **28**(1), 85–89 (2007)

K. Okabe, T. Koshinaka, K. Shinoda, Attentive Statistics Pooling for Deep Speaker Embedding, in *Proc. INTERSPEECH*, 2018, pp. 2252–2256

M. Pal, D. Paul, G. Saha, Synthetic speech detection using fundamental frequency variation and spectral features. Computer Speech & Language **48**, 31–50 (2018)

T.B. Patel, H.A. Patil, Cochlear filter and instantaneous frequency based features for spoofed speech detection. IEEE Journal of Selected Topics in Signal Processing **11**(4), 618–631 (2017)

D. Paul, M. Pal, G. Saha, Spectral features for synthetic speech detection. IEEE Journal of Selected Topics in Signal Processing **11**(4), 605–617 (2017)

D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing **3**(1), 72–83 (1995)

M. Sahidullah, T. Kinnunen, C. Hanilçi, A comparison of features for synthetic speech detection, in *Proc. INTERSPEECH*, 2015, pp. 2087–2091

M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, K.-A. Lee, Introduction to Voice Presentation Attack Detection and Recent Advances, ed. by S. Marcel, M.S. Nixon, J. Fierrez, N. Evans (Springer, Cham, 2019), pp. 321–361

X. Tian, X. Xiao, E.S. Chng, H. Li, Spoofing speech detection using temporal convolutional neural network, in *ASIPA*, 2016, pp. 1–6

M. Todisco, H. Delgado, N. Evans, Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language **45**, 516–535

(2017)
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. arXiv preprint arXiv:1904.05441 (2019)
F. Tom, M. Jain, P. Dey, End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention, in *Proc. INTERSPEECH*, 2018, pp. 681–685
T.M. Ventura, A.G. de Oliveira, T.D. Ganchev, J.M. de Figueiredo, O. Jahn, M.I. Marques, K.-L. Schuchmann, Audio parameterization with robust frame selection for improved bird identification. Expert Systems with Applications **42**(22), 8463–8471 (2015)
J.A. Villalba, A. Miguel, A. Ortega, E. Lleida, Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge, in *Proc. INTERSPEECH*, 2015a, pp. 2067–2071
J. Villalba, A. Miguel, A. Ortega, E. Lleida, Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge, in *Proc. INTERSPEECH*, 2015b
Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge, in *Proc. INTERSPEECH*, 2015. 2037–2041
Z. Wu, et al., ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. IEEE Journal of Selected Topics in Signal Processing **11**(4), 588–604 (2017)
Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: A survey. Speech Communication **66**, 130–153 (2015)
H. Yu, Z.-H. Tan, Z. Ma, R. Martin, J. Guo, Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features. IEEE transactions on neural networks and learning systems **29**(10), 4633–4644 (2017)
Y. Zhu, T. Ko, D. Snyder, B. Mak, D. Povey, Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification, in *Proc. INTERSPEECH*, 2018, pp. 3573–3577