Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 1, 2023, 119-129.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection

Pedram Abdzadeh Ziabari and Hadi Veisi*

Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran.

Article Info

Article History:
Received 28 October 2022
Revised 20 November 2022
Accepted 18 December 2022

DOI:10.22044/jadm.2022.12373.2382

Keywords:
Voice Spoofing Detection, Deep Neural Networks, Voice Biometrics, Deepfake Audio Detection.

**Corresponding author: h.veisi@ut.ac.ir (H. Veisi).*

Abstract

Automatic Speaker Verification (ASV) systems have proven to be vulnerable to various types of presentation attacks, among which Logical Access (LA) attacks are manufactured using voice conversion and text-to-speech methods. In the recent years, there has been loads of work concentrating on synthetic speech detection, and with the emergence of deep learning-based approaches and their success in a variety of computer science fields, they have been a prevailing tool for this very task too. Most of the deep neural network-based techniques for synthetic speech detection have employed acoustic features based on Short-Term Fourier Transform (STFT), which are extracted from the raw audio signal. However, lately, it has been discovered that the usage of Constant Q Transform's (CQT) spectrogram can be a beneficial asset both for performance improvement and processing power and time reduction of a deep learning-based synthetic speech detection. In this work, we provide a comparison between the usage of the CQT spectrogram and some most utilized STFT-based acoustic features. Utilization of a ResNet-based architecture contemplated in this work as this architecture has had lots of success in deepfake speech detection. As lateral objectives, we consider improving the performance of the model as much as we can using methods such as self-attention and one-class learning. Also short-duration synthetic speech detection has been one of the lateral goals too. Finally, we see that the CQT spectrogram-based model not only outperforms the STFT-based acoustic feature extraction methods but also reduces the processing time and resources for the detection of genuine speech from the fake. Also the CQT spectrogram-based model places well among the best works done on the LA subset of the ASVspoof 2019 dataset, especially in terms of Equal Error Rate (EER).

1. Introduction

Biometrics have been of paramount importance for the identification of the users of systems, and have the ability to remove the need for the rigorous task of remembering usernames and passwords for various systems, websites, and mobile applications. Face, iris, voice, and fingerprint of the user trying to enter the main system, are the main biometrics that can be utilized for user identification and verification. Automatic Speaker Verification (ASV) systems [1], systems that have the primary objective of verifying the identity of a user trying to enter the main system, have been fairly attractive; hence lots of research have been done on them. However, similar to systems based on other biometrics, it is known that ASV systems are susceptible to presentation attacks. Presentation of an artifact or made-up human characteristic to a biometric system to circumvent security and abuse the system is known as a presentation attack. Text-To-Speech (TTS) [2], Impersonation [3], Voice Conversion (VC) [4], and replay [5] are the

predominant voice presentation attacks used to penetrate ASV systems. When a fraudulent speech signal is directly injected into the ASV system, a Logical Access (LA) attack occurs. On the other hand, when the signal goes through a microphone, the attack is called a Physical Access (PA) attack [6]. Figure 1 recapitulates voice spoofing attacks. To immune ASV systems against voice presentation attacks, various voice spoofing detection systems are created. In the recent years, and with the increase in the usage of ASV systems, the attention to voice spoofing detection has also been rising up. In 2015, the ASVspoof challenge [7] was inaugurated to address the voice spoofing detection problem, provide benchmarks and datasets, and consequently, boost voice spoofing countermeasures' capability to discern presentation attacks. The 2015 version of the ASVspoof challenge was mainly concentrated on synthetic speech detection, while The 2017 version was focused on the replay attack. However, the ASVspoof 2019 contemplated both LA and PA attacks. For the creation of the LA subset of the ASVspoof 2019 database, the state-of-the-art methods of VC and TTS were exploited [8]. The ASVspoof 2021 [9] has provided a further enriched evaluation subset of the database in terms of variety and quality of attacks, and also, the challenge had a section dedicated to the detection of compact deepfake audio files.

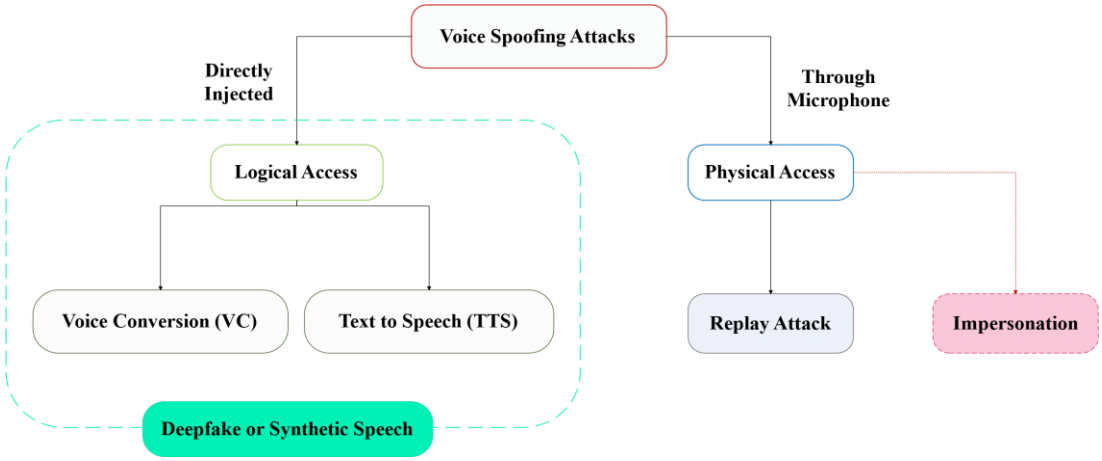


Figure 1. A review of different types of voice spoofing attacks [10].

Loads of different methods have been applied to detect synthetic speech during the years of research in this field. The very first works on synthetic speech detection were mainly concentrated on feature engineering and the suggestion of novel and more beneficial acoustic features while using classifiers such as Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). With the emergence of deep learning-based models and their vast success all over the computer science fields, from computer vision to natural language processing, and with the ability to be applied for different tasks such as classification and regression, it is not unexpected that they have employed in most of the recent synthetic speech detection works. In fact, Deep Neural Networks (DNN)-based approaches have been preponderating in voice spoofing detection [11]–[14]. Mainly, acoustic features resulting from Fourier analysis such as spectrogram of Short Term Fourier Transform (STFT), Linear Frequency Cepstral Coefficients (LFCC), and Mel-Frequency Cepstral Coefficients (MFCC) extracted from the speech signal have been contemplated as the input of various DNNs for voice spoofing detection and speech processing in general [15], [16]. However, recently, Constant Q Transform (CQT), a transform initially inaugurated with music processing objectives [17] has shown capabilities to be applied in other audio processing realms, such as audio signal classification and separation. The Constant Q Cepstral Coefficient (CQCC) features [18] driven from cepstral analysis of CQT were the first CQT-related features considered in works in voice spoofing detection [11], [13]. However, in the recent works, extraction of the CQT spectrogram from the speech and its usage of it as the acoustic features that are given as input to the network has proven to be lucrative in terms of processing time reduction [10] and performance improvement [19]. In fact, in [19], the state-of-the-art result on synthetic speech detection on the evaluation subset of the LA subset of the ASVspoof 2019 challenge has been achieved while using the CQT spectrogram as input of the DNN.

In this work, building upon our work in [10], we pursue the goal of comparing the newly coming into attention acoustic feature CQT spectrogram with well-established Fourier transform-based

acoustic features to find more about the capabilities of this feature. Even though current methods of synthetic speech detection have been predominantly concentrated on the neural network architecture and proposing inventions in this regard, the input of the neural network still possesses enough importance to be talked about and can highly influence the outcome of the model. Hence, our work can show how each one of the CQT spectrogram and STFT-based features can affect the performance of a deepfake speech detection model and the required processing power and time for it. As the aforementioned factors are the main ones in the efficiency and performance analysis of a model. Hence, this comparison can help with future research choices of acoustic features to use with their DNN-based model. For the comparison of the STFT-based features and CQT spectrogram, we first extract the CQT spectrogram, STFT spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), and linear frequency cepstral coefficients for the raw speech signal. Then we feed each feature as input to a Resnet-18 model, which has been empowered with a temporal self-attention layer and a feedforward partition that is created from three multi-perceptron layers. In the end, we use one class learning proposed in [20] for score and loss calculation and better generalization against unseen attacks. The reason behind the usage of a ResNet architecture, other than the success of these models in computer vision since their inauguration, is their popularity and usage of them in many works in the deepfake speech detection field [10], [11], [13], [20]–[22]. Hence, most likely, a ResNet-based model can help us reach a more general conclusion for the comparison of features in deep learning-based synthetic speech detection.

As a lateral goal, we want to use the model to reach the best possible result to raise the performance in the detection of synthetic speech the most. The reason is that even though the number of works concentrating on synthetic speech detection is growing in recent years, at the same time, the number of novel spoofing attacks and ways of creating synthetic speeches is growing, too, even in a quicker way. Hence we are using different mechanisms, such as one-class learning to improve the performance against unseen attacks [20]. Also we are using the attention mechanism, which has had lots of success in natural language processing. As another lateral factor, we are considering the employment of short-duration utterances for the task of deepfake speech detection, as the shorter the length of the input signal is, the better the user experience can get working with an ASV system. As a result, recently, more research has been focusing on Short-duration Speaker Verification (SdSV) tasks [23]–[26]. The reason is the shorter the After implementing the models based on each acoustic feature, and it is observable that CQT spectrogram-based model reaches the best EER and min t-DCF among all the features mentioned earlier, while it uses noticeably less processing power and requires less time for training and test. Moreover, the CQT spectrogram-based model places among the best performers in the field of synthetic speech detection, while a considerably shorter utterance is being used compared to the state-of-the-art models.

The rest of this paper is structured as what follows. Some related works are reviewed in Section 2. Section 3 gives some insight into the background needed for a more vivid understanding of this paper. Section 4 elaborates on the details of the proposed method, and Section 5 describes the experimental setup and results. Finally, in Section 6, the work done in this paper is concluded.

2. Related Works

In this section, the works done in synthetic speech detection are reviewed briefly. We divide works in this field into Traditional Approaches, which are classic works done before the emergence of deep learning (mostly based on signal processing and feature engineering), and Deep Learning-based approaches.

2.1. Traditional approaches

Early works on synthetic speech detection were concentrated on feature discovery [27]. They mostly have considered the combination of the Support Vector Machine (SVM) or Gaussian Mixture Model (GMM) (for the classification of genuine and counterfeit inputs) with various speech representations, such as LFCC, MFCC, and magnitude and phase spectrum of the speech signal. Wu *et al*. [28] have employed a GMM with MFCC, modulation features, and modified group delay cepstral coefficients as input features. In [29], various acoustic features utilized for fake speech detection have been compared with the usage of the GMM and the SVM as classifiers. Also In [30], Phase-information-related features such as Relative Phase Shift (RPS) and Modified Group Delay (MGD) were exercised with GMM for classification. For the first time in [31], Todisco *et al*. gave CQCC features extracted from utterances to a GMM classifier to separate fake from authentic speech.

2.2. Deep learning-based approaches

Recently and with the astonishing results of deep learning-based models in different computer science areas, the contributions in synthetic speech and voice spoofing detection have been dominated by end-to-end deep learning methods. In [11], a fusion of ResNets has been considered with CQCC, MFCC, and log magnitude of STFT spectrogram features of the input signal for the tasks of deepfake speech and replay attack detection. Also Gomez-Alanis *et al*. [12] have utilized a Light Convolutional Neural Network (LCNN) to extract features and a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU) to learn the long-term dependencies of audio signals. In [13], different configurations of ResNets and squeeze-excitation networks were used to create a countermeasure. Chettri *et al*. [32] employed deep models such as CNN and CRNN along with shallow models constructed of GMM and SVM to create three ensemble models for synthetic speech detection. Usage of Acoustic features of utterances such as Static, Delta, and Acceleration (SDA), MFCC, IMFCC, CQCC, and Sub-band Centroid Magnitude Coefficients (SCMC) was considered in their work. In [14], it is noted that genuine speech samples have relatively low variance in characteristics compared with synthetic speech, and a model based on LCNN and feature genuinization is proposed. Monteiro *et al*. [21] used the modified group delay function feature and STFT spectrogram along with LCNN and ResNet equipped with temporal self-attention. A combination of graph attention networks and ResNet-18 were used in [22], with Log-linear Filter-Bank (LFB) extracted from utterances as input to distinguish synthetic speech from real samples. A Resnet-18 with self-attention and one-class learning has been trained on LFCC features in [20]. In [33], Fang *et al*. have proposed a Dual Path Res2Net (DP-Res2Net) model, which takes the raw waveform as its input. Also with the generalizability and improvement in mind as the main goal, the [34] authors have considered the exploitation of prototypical loss under the meta-learning paradigm and have utilized Squeeze Excitation Residual Network (SE-ResNet). Ma *et al*. [35] have used a knowledge distillation-based loss function and have applied continual learning for the first time in synthetic speech detection. In [10], a countermeasure based on the CQT spectrogram has been proposed, with a ResNet-18 at its core and temporal self-attention. Li et al. [19] have proposed a channel-wise gated Res2Net model adopting the CQT spectrogram as the acoustic feature and have reached the state-of-the-art results on the ASVspoof 2019 LA dataset. A summary of works on deep learning-based synthetic speech detection is available in Table 1.

Table 1. Summary of works in Deep Learning-based synthetic speech detection.

Year	Author(s) and Reference	Method	Dataset	EER (%)	Min t-DCF
2019	Alzantot *et al*. [11]	Fusion of (CQCC,MFCC, Spectrogram) + ResNet	ASVspoof 2019	6.02	0.156
2019	Gomez-Alanis *et al*. [12]	Spectrogram + LCNN + GRU-RNN	ASVspoof 2019	6.28	0.152
2019	Lai *et al*. [13]	Fusion of (Log Spectrogram, CQCC) + (SENet, ResNet)	ASVspoof 2019	6.70	0.155
2019	Chettri *et al*. [32]	Ensemble of (Log-spec, Log-Mel-spec) + (CNN, CRNN, Wave-U-Net, …)	ASVspoof 2019	2.64	0.075
2020	Wu *et al*. [14]	Log Power Spectrogram + LCNN with Feature Genuinization	ASVspoof 2019	4.07	0.102
2020	Monteiro *et al*. [21]	(LFCC+Product Spectrum) + (LCNN, ResNet)+Self-attention	ASVspoof 2019	9.87	0.1890
2021	Tak *et al*. [22]	LFCC+ ResNet+Graph Attention Network	ASVspoof 2019	1.68	0.047
2021	Zhang *et al*. [20]	LFCC + ResNet + Self-attention + One Class Learning	ASVspoof 2019	2.19	0.059
2021	Ziabary and Veisi [10]	CQT Spectrogram + ResNet + Self-attention + One Class Learning	ASVspoof 2019	3.53	0.10
2021	Li *et al*. [19]	CQT Spectrogram +Multi-group Channel-wise Res2Net	ASVspoof 2019	1.78	0.052
2021	Fang *et al*. [33]	Waveform + Dual Path Res2Net + AM-Softmax	ASVspoof 2019	0.47	-
2021	Ma *et al*. [35]	Continual Learning	ASVspoof 2019	7.74	-
2022	Pal *et al*. [34]	Meta-Learning + (SE-ResNet, ResNet) + (Prototypical Loss, One Class Softmax)	ASVspoof 2019	2.00	0.048

3. Background

3.1. Short-time Fourier transform (STFT)

Short-time Fourier Transform (STFT) is an extension of the Fourier transform, and consists of a series of Fourier transform of a windowed signal, computed every t milliseconds. By this method, a signal is decomposed into a series of short segments, a
nd short-term frequency information for each frame of the signal is extracted.

The short-time Fourier transform of the mth frame of a speech signal is calculated as:

$$X_m(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x_m[n]e^{(-j\omega n)} \quad (1)$$

where:

$$x_m[n] = w[m-n]x[n] \quad (2)$$

and w is the window function which, except in a small region, possesses a value of zero.
The STFT is normally visualized by means of its spectrogram, which is the intensity plot of STFT magnitude over time. An example of an STFT spectrogram is shown in Figure 2.

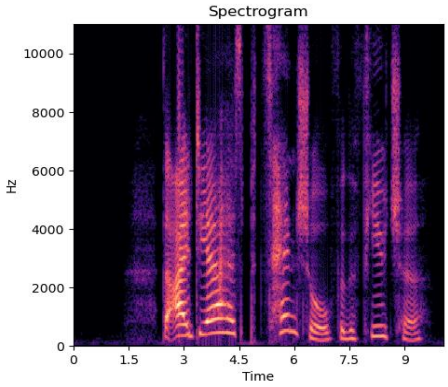


Figure 2. An illustration of an example STFT spectrogram

3.2. Linear frequency and Mel-frequency cepstral coefficients (LFCC and MFCC)

Filter bank-based cepstral features such as Linear Frequency Cepstral Coefficients (LFCC) and Mel-Frequency Cepstral Coefficients (MFCC) have been contemplated immensely as a way for acoustic signal analysis. These features have been inaugurated to overcome the high-dimensionality issues of acoustic features such as the STFT spectrogram.

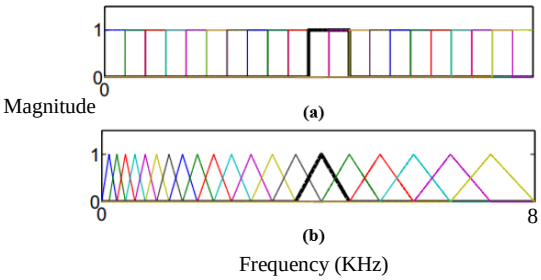


Figure 3. a) Linear filters for LFCC b) Mel-scale filters for MFCC [29].

To extract the MFCC and LFCC features, first, overlapping band-pass filters integrate the power spectrum of an audio signal (or each frame of an audio signal). Then there will be a logarithmic compression step, and afterward, the Discrete Cosine Transform (DCT) is performed to produce cepstral coefficients. For the production process of LFCC features, filters have a triangular shape. For the MFCC, however, the overlapping triangular filters are placed in the Mel scale. The difference between filters used in the LFCC and the MFCC is visible in Figure 3.

3.3. Constant Q transform

The Constant Q Transform (CQT) was inaugurated in 1991 [17] mainly aiming for music processing objectives. The CQT provides a constant Q factor (the ratio between the center frequency f_k and the bandwidth $\delta f, Q = f_k / \delta f$) all over the spectrum, contradictory to the Short-Term Fourier Transform (STFT) that is mainly employed for the computation of a speech signal's spectrogram and has a changing Q factor. Moreover, the STFT spectrogram lacks time resolution at higher frequencies and lacks frequency resolution at lower frequencies, while CQT benefits from higher time resolution at higher frequencies in addition to higher frequency resolution and low frequencies [31]. These abilities are achieved by the way of computation of the CQT. The better frequency resolution in lower frequencies is observable in lower parts of Figure 4 (b) where the distance between the CQT spectrograms points is very short, while more time resolution in higher frequencies is clear in the higher parts of the CQT spectrogram. Moreover, the research in [36] shows that usage of the CQT in shifted non-negative matrix factorization models can enhance the quality of individual sound source separation. Additionally, in [37], for acoustic scene classification, a fusion of the STFT spectrogram and the CQT spectrogram features and CNNs was suggested. It was shown that the fused model could reach better results than the solitary STFT spectrogram model. Moreover, as

mentioned earlier, in [19], the CQT spectrogram has been used along with a multi-group channel-wise Res2Net which has achieved results close to ones of the state-of-the-art model.

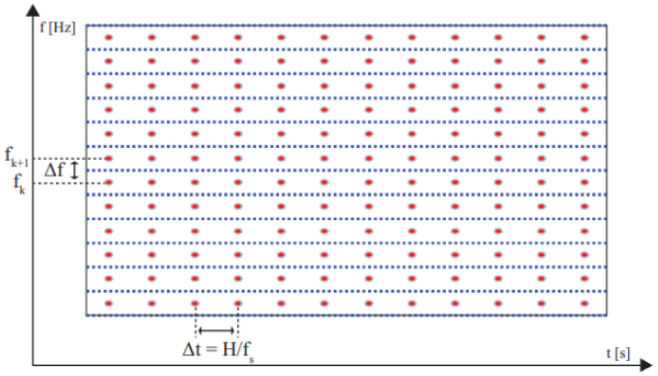
3.3.1. CQT computation

According to [17], to compute the CQT of a time-domain signal $x(n)$, the following formulation is used:

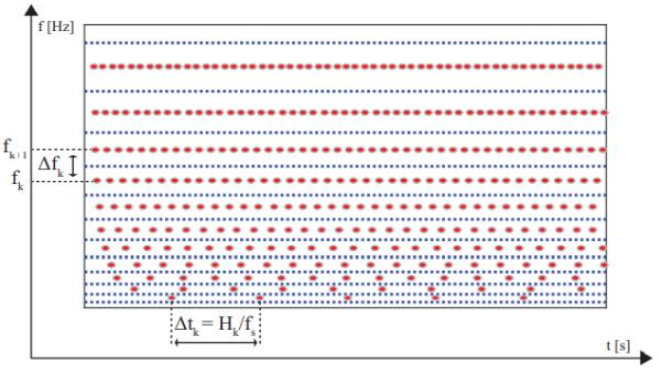
$$X^{CQ}(k,n)=\sum_{j=n-\left\lfloor \frac{N_k}{2} \right\rfloor}^{n+\left\lfloor \frac{N_k}{2} \right\rfloor} x(j)a_k^*(j-n+\frac{N_k}{2}) \quad (3)$$

where N_k is the kth window length, and $k=1,2,\dots,K$ is the frequency bin index. a_k^* is the complex conjugate of a_k and a time-frequency atom $a_k(n)$ is calculated according to:

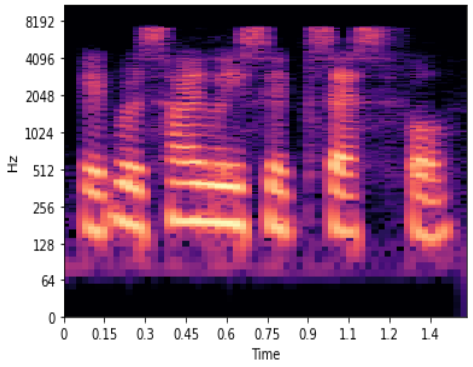
$$a_k(n)=\frac{1}{C}(\frac{n}{N_k})\exp\left[i\left(2\pi n\frac{f_k}{f_s}+\Phi_k\right)\right] \quad (4)$$



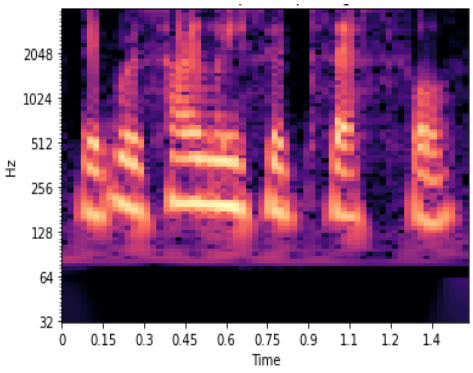
(a) STFT



(b) CQT



(c) STFT for a Sample Audio



(d) CQT for a Sample Audio

Figure 4. A time-frequency domain comparison of a) STFT and b) CQT spectrograms. H is the duration for sliding window analysis [31] c) STFT spectrogram for a sample audio, d) CQT spectrogram for the same sample.

where f_s indicates the sampling rate and f_k is the center frequency of the kth bin. Also Φ_k is the phase offset and the scaling factor C calculated through:

$$C=\sum_{l=-\left\lfloor \frac{N_k}{2} \right\rfloor}^{\left\lfloor \frac{N_k}{2} \right\rfloor} w\left(\frac{l+\frac{N_k}{2}}{N_k}\right) \quad (5)$$

Additionally, the f_k in (2) obeys the following formula:

$$f_k=f_1 2^{\frac{k-1}{B}} \quad (6)$$

where B is the number of bins per octave and f_1 shows the center frequency for the lowest frequency bin. Also the following equation gives the Q factor:

$$Q = \frac{f_k}{(f_{(k+1)} - f_k)} = (2^{(\frac{1}{B})} - 1)^{(-1)} \quad (7)$$

Finally, the window length N_k is:

$$N_k = \frac{f_s}{f_k} Q \quad (8)$$

4. Proposed Method

In this section, we give the details about the used acoustic features of a speech signal and the method we use for comparison of them in deep learning-based synthetic speech detection. Figure 5 gives an overview of our model.

4.1. Input

In the first step, different acoustic features are extracted from the raw signal using the methods described in Section 3. The STFT spectrogram, LFCC features per each frame of an audio signal, MFCC features per each frame of an audio signal, and CQT power spectrogram are separately given as inputs to the neural network architecture discussed in the following paragraphs.

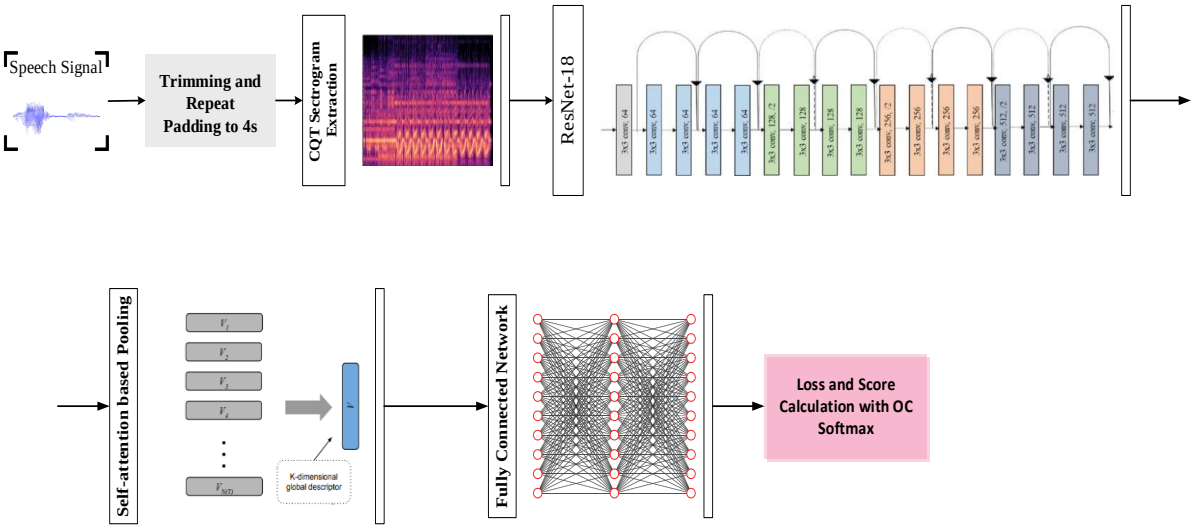


Figure 5. An overview of all the steps of our model.

4.2. Neural network model

A ResNet-18 model has been contemplated as the backbone for feature extraction from the input. The immense success of ResNet-based models is the primary reason for the selection of such a network (In [21], it is shown that larger Residual Networks such as ResNet-50 are more prone to overfitting in the task of synthetic speech detection on the ASVspoof 2019 dataset), because of the higher number of parameters in ResNet-50 compared with ResNet-18, especially when combined with attention layers. The representations resulting from the application of ResNet-18 are then mapped to a set of vectors. Afterward, based on the prevalence of attention-based mechanisms in various tasks such as natural language processing, there is a self-attention layer that gets the aforementioned set of vectors and outputs a new vector that contains the temporal importance of each part of the input. In the end, we have a feedforward network, which is consisted of three fully connected layers.

4.3. Loss function

While the characteristics of fake audio created by different types of attacks can be disparate from one attack to another, there is much more similarity in the attributes of samples of genuine speech. Hence, we consider the usage of the one-class learning approach presented in [20], where the One-Class Softmax (OCS) loss function is calculated through:

$$L = \frac{1}{N} \sum_{(i=1)}^{N} log\left(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \hat{x}_i)(-1)^{y_i}}\right) \quad (9)$$

where w_0 is the only weight vector, x is the input vector, and y denotes the target label. Also m and N are the cosine similarity margin and the number of samples in a batch, respectively. Finally, α is a scale factor, and $\hat{w}$ and $\hat{x}$ are normalized versions of w and x. In [20], it is shown that OCS loss can attain satisfactory results and can further enhance the ability of the model to discern unknown attacks from genuine speech.

5. Experiments

In this section, we elaborate on the practical details of the work done in this paper.

5.1. Dataset

The LA subset of the well-known ASVspoof 2019 challenge dataset [8] that has established itself as a benchmark in the area of voice spoofing detection and synthetic speech detection is used for the training and evaluation of the variation of the proposed model. Various Text to Speech (TTS) and Voice Conversion (VC) methods have been employed to create fake speeches in the LA subset of the ASVspoof 2019 dataset. The LA subset is divided into Train, Development, and Evaluation parts. We have merged the training and development parts in our work and used the resulting section in the training and validation process. The same six different types of synthetic speech creation techniques are utilized in both the Training and Development subsets of LA. However, the evaluation set is comprised of 11 types of attacks previously unseen in the other sets. Table 2. elaborates on some details about the LA subset of the ASVspoof 2019 dataset.

Table 2. Details of LA subset of ASVspoof 2019 challenge dataset.

	#Speakers		#Utterances		Duration
Subset	Female	Male	Spoof	Bonafide	
Training	12	8	22800	2580	24h:10m
Development	12	8	22296	2548	24h:15m
Evaluation	-	-	71747		62h:44m

5.2. Evaluation metrics

Equal Error Rate (EER) and minimum tandem detection cost function (min t-DCF) metrics have been deployed to evaluate and compare the proposed methods. These are the two prevailing metrics used for the comparison of different voice spoofing detection systems in previous research works. The min t-DCF metric was introduced at the ASVSpoof 2019 challenge, aiming to assess the performance of an anti-spoofing countermeasure in the presence of an ASV system and determine how much a misclassification in countermeasure affects the ASV system. On the other hand, EER is the threshold where the false alarm rate and miss rate are the same and is an already established metric in the evaluation of the performance of biometrics and anti-spoofing systems. Lower values in terms of both EER and min t-DCF show higher degrees of performance of the countermeasure.

5.3. Training details

In the training phase for all models, first, trimming and repeat padding are used to create input utterances of equal length of 4 seconds, and then the inputs are normalized. After that, each feature is extracted from the raw input signals for each model, respectively. For the MFCC and LFCC features, 20 frequency filters are considered for each frame of the data, and the delta and delta-delta of features are combined with them. As a result, the extracted feature vector for MFCC and LFCC has a 60*251 size. For the STFT spectrogram, 512 frequency bins have been considered for each frame of data, with rectangular overlapping windows, and a hop length of 128, and at the end, the resulting vector is 257*251. The GPU implementation of CQT proposed in [23] has been employed for the creation of CQT spectrogram-based features, with a sampling rate of 16KHz, 84 frequency bins, a hop length of 512 samples, 12 bins per octave, and 32.70Hz as the frequency for the lowest CQT bin. This configuration results in a size of 84*126 for the power CQT spectrogram of each utterance as the input.

As stated earlier, each one of the extracted acoustic features is given to a combination of the ResNet-18 with self-attention, and feedforward layers and the one-class softmax layer proposed in [20] are employed for the implementation of models. At the end of these layers, an embedding vector with a size of 256 is extracted. Finally, the embedding vector passes through the one-class softmax layer for the calculation of loss and score. Additionally, the same values as [20] have been considered for the loss function parameters; $\alpha = 20$ and $m = 0.9$. We trained all models for more than 100 epochs on the Kaggle.com servers.

5.4. Results and Comparison

After computing the scores for each utterance of the test set and calculating EER and min t-DCF for each model, it is observable that for the task of deepfake speech detection on the LA subset of the ASVspoof 2019 dataset, the model with CQT spectrogram as a feature extracted from each utterance, has the best results compared with the models based on other acoustic features. The CQT power spectrogram-based model reaches an EER

of 2.33% and min t-DCF 0.12, which improves baseline models by about 71% in terms of EER. As it can be seen in Table 3, CQT based model not only outperforms models based on LFCC, MFCC, and STFT spectrogram on the grounds of performance but also, as a result of smaller feature vector size, it needs the least amount of computation time and graphical memory, so that the STFT spectrogram based model needs 300% more graphical memory and about three times of training duration per epoch.

As it was mentioned in Section 5.3, the parameters considered for extraction of CQT spectrogram features result in smaller feature vectors for the CQT spectrogram than other STFT-based features. Hence, the usage of CQT-based features not only has improved the performance but also has reduced the required processing power drastically. Consequently, the user experience of the biometric system will be improved, and the implementation cost of the system will be less compared to when other features are being contemplated.

Table 3. Comparison of acoustic features when combined with ResNet-18+Attention+MLP+OC learning.

Acoustic feature	EER	Min t-DCF	Avg. Time / Epoch	Approx. Max GPU RAM
LFCC	4.7	0.15	245s	4.5GB
MFCC	8.33	0.223	225s	4.5GB
STFT Spectrogram	5.05	0.186	600s	10GB
CQT power spectrogram	**2.33**	**0.120**	**140s**	**2.4GB**

Table 4. Comparison of CQT Model with some of the recent works

* The input dimesion is not calculated as the model has a weaker performance than proposed model

System	EER	min t-DCF	Input Dim
CQCC + GMM-baseline [8]	9.57	0.257	*
LFCC + GMM-baseline [8]	8.09	0.212	*
Lai *et al*. [13]	6.70	0.155	*
Alzantot *et al*. [11]	6.02	0.156	*
Wu *et al*. [14]	4.07	0.102	*
Proposed (CQT power spectrogram)	**2.33**	**0.120**	**84*126**
Zhang *et al*. [20]	2.19	0.059	60*1498
Pal *et al*. [34]	1.70	0.048	60*1498
Tak *et al*. [22]	1.68	0.047	60*404

Also CQT-based model places among top performers even though it is lighter than most models in terms of input dimensions, and any kind of model fusion or data augmentation has not been used, unlike some other works. Table 4 compares the proposed models with ASVspoof 2019 baselines and some of the recent works in synthetic speech detection. In Table 4, the input dimensions of works that are performing better than the proposed method, are calculated and as it is evident they have all higher input dimensions.

5.5. On the effects of short-duration of utterances

As it was spoken of in the previous sections, utterances of length 4 seconds have been used for feature extraction and synthetic speech detection. When compared to other works such as [20], which uses utterances of the length of 15 seconds, we are clearly using a shorter duration of utterances for deepfake speech detection. Short-duration speaker verification has become a trend in recent years [24]–[26]. The Short-duration Speaker Verification (SdSV) Challenge has been created focusing on this subject, aiming to speaker verification in more realistic situations. Synthetic speech detection in a shorter duration, which is observable in our work, can help the combination of ASV systems and countermeasure work to achieve a more convenient user experience (as the user needs to speak less) and also need comparably less amount of processing power and computational resources.

6. Summary and Conclusion

In this work, we provided a comparison between the STFT-based acoustic features and the CQT spectrogram in terms of repercussions of their usage in deep learning-based synthetic speech detection. It is observable that when used with our ResNet-based DNN architecture, not only does the CQT spectrogram proves to be more effective performance-wise, but also, in terms of required time and processing power needed, it tops STFT-based acoustic features. In the future work, other network architectures such as transformers and Res2Nets can be used for this comparison to give more insight into the difference between these features. Also, we have not dedicated much time to the analysis of the difference that variations of parameters of CQT could make, which is the work that can be done in the future.

References

[1] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, no. 1, pp. 91–108, Aug. 1995.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[3] Yee Wah Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, Oct. 2004, pp. 145–148.

[4] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Nov. 2013, pp. 1–9.

[5] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec. 2014, pp. 1–5.

[6] M. Todisco *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *ArXiv190405441 Cs Eess*, Apr. 2019.

[7] Z. Wu *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," presented at the Sixteenth annual conference of the international speech communication association, 2015.

[8] J. Yamagishi *et al.*, "Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database," 2019.

[9] J. Yamagishi *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," *ArXiv210900537 Cs Eess*, Sep. 2021

[10] P. A. Ziabary and H. Veisi, "A Countermeasure Based on CQT Spectrogram for Deepfake Speech Detection," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, Dec. 2021, pp. 1–5.

[11] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Interspeech 2019*, Sep. 2019, pp. 1078–1082. doi: 10.21437/Interspeech.2019-3174.

[12] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," in *Interspeech* 2019, Sep. 2019, pp. 1068–1072.

[13] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *Interspeech 2019*, Sep. 2019, pp. 1013–1017.

[14] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks," in *Interspeech 2020*, Oct. 2020, pp. 1101–1105.

[15] K. Aghajani, "Audio-visual emotion recognition based on a deep convolutional neural network," *Journal of AI & Data Mining*, vol. 10, no. 4, pp. 529–537, Nov. 2022.

[16] B. Z. Mansouri, H. R. Ghaffary, and A. Harimi, "Speech Emotion Recognition using Enriched Spectrogram and Deep Convolutional Neural Network Transfer Learning," *J. AI Data Min.*, vol. 10, no. 4, pp. 539–547, Nov. 2022.

[17] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[18] M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," in *The Speaker and Language Recognition Workshop (Odyssey 2016)*, Jun. 2016, pp. 283–290.

[19] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks," *ArXiv210708803 Cs Eess*, Jul. 2021, Accessed: May 02, 2022. [Online]. Available: http://arxiv.org/abs/2107.08803

[20] Y. Zhang, F. Jiang, and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[21] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, p. 101096, Sep. 2020.

[22] H. Tak, J. Jung, J. Patino, M. Todisco, and N. Evans, "Graph attention networks for anti-spoofing," *ArXiv Prepr*. ArXiv210403654, 2021.

[23] Z. Huang, S. Wang, and K. Yu, "Angular Softmax for Short-Duration Text-independent Speaker Verification.," presented at the Interspeech, 2018, pp. 3623–3627.

[24] M. Sahidullah *et al.*, "UIAI System for Short-Duration Speaker Verification Challenge 2020," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 323–329.

[25] S. Wang, Z. Huang, Y. Qian, and K. Yu, "Discriminative Neural Embedding Learning for Short-Duration Text-Independent Speaker Verification," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1686–1696, Nov. 2019.

[26] Y. Jung, Y. Choi, H. Lim, and H. Kim, "A Unified Deep Learning Framework for Short-Duration Speaker Verification in Adverse Environments," *IEEE Access*, vol. 8, pp. 175448–175466, 2020.

[27] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Trans. Signal Inf. Process.*, vol. 9, p. e2, 2020.

[28] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7234–7238.

[29] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A Comparison of Features for Synthetic Speech Detection," 2015, Accessed: May 02, 2022. [Online]. Available: https://erepo.uef.fi/handle/123456789/4371

[30] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Phase-Aware Signal Process. Speech Commun.*, vol. 81, pp. 30–41, Jul. 2016.

[31] M. Todisco, H. Delgado, and N. W. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients.," presented at the Odyssey, 2016, vol. 2016, pp. 283–290.

[32] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble Models for Spoofing Detection in Automatic Speaker Verification," in *Interspeech 2019*, Sep. 2019, pp. 1018–1022.

[33] X. Fang, H. Du, T. Gao, L. Zou, and Z. Ling, "Voice Spoofing Detection with Raw Waveform Based on Dual Path Res2net," in *5th International Conference on Crowd Science and Engineering*, New York, NY, USA, 2021, pp. 160–165.

[34] M. Pal, A. Raikar, A. Panda, and S. K. Kopparapu, "Synthetic speech detection using meta-learning with prototypical loss," *ArXiv220109470 Cs Eess*, Jan. 2022.

[35] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual Learning for Fake Audio Detection," *ArXiv210407286 Cs Eess*, Apr. 2021.

[36] R. Jaiswal, D. Fitzgerald, E. Coyle, and S. Rickard, "Towards Shifted NMF for Improved Monaural Separation," *IET Conf. Proc.*, pp. 19-19(1), Jan. 2013.

[37] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," *Detect. Classif. Acoust. Scenes Events DCASE*, 2017.

مقایسه طیف CQT با ویژگیهای آکوستیک مبتنی بر STFT در تشخیص گفتار ساختگی مبتنی بر یادگیری عمیق

پدرام عبدزاده ضیابری و هادی ویسی*

گروه بین رشته‌ای فناوری دانشکده علوم و فنون نوین، دانشگاه تهران، تهران ، ایران.

ارسال ۲۰۲۲/۱۰/۲۸؛ بازنگری ۲۰۲۲/۱۱/۲۰؛ پذیرش ۲۰۲۲/۱۲/۱۸

چکیده:

سیستم‌های تایید صحت گوینده نسبت به حملات ارائه صوت آسیب پذیر هستند، که در میان آنها حملات دسترسی منطقی با استفاده از ابزارهایی مانند تبدیل صوت و متن-به-صوت ایجاد می شوند. در سال‌های اخیر، پژوهش‌های زیادی متمرکز بر شناسایی گفتار ساختگی بوده‌اند و با ظهور الگوریتم‌های یادگیری عمیق و موفقیت آن‌ها در علوم کامپیوتر، الگوریتم غالب در این حوزه نیز به شمار می‌روند. اکثر روش‌های مبتنی بر یادگیری عمیق برای شناسایی گفتار ساختگی از ویژگی‌های مبتنی برSTFT که از سیگنال اولیه استخراج می‌شوند استفاده کرده‌اند. در حالی که اخیرا مشخص شده است که استفاده از طیف CQT می‌تواند ابزار مناسبی برای بهبود عملکرد و استفاده بهینه از توان پردازشی و کاهش زمان پردازش باشد. در این پژوهش، مقایسه ای بین استفاده از طیف CQT و چندین روش مبتنی بر STFT که بیشترین استفاده را دارند ارائه می‌شود. در این کار از یک معماری مبتنی بر ResNet استفاده شده است، چرا که این معماری به موفقیت‌های فراوانی در در زمینه تشخیص گفتار جعلی دست یافته است. به عنوان هدف جانبی، ما به بهبود حداکثری مدل ارائه شده با استفاده از ابزارهایی مانند یادگیری تک کلاسه و توجه-به-خود می‌پردازیم. همچنین شناسایی گفتار ساختگی با طول جملات کوتاه از سایر اهداف مورد توجه در کار ماست. در نهایت، مشخص می‌شود که مدل مبتنی بر طیفCQT از لحاظ عملکرد، منابع و زمان پردازش بهتر از مدل های مبتنی بر STFT عمل می‌کند. همچنین، مدل مبتنی برCQT در جایگاه مناسبی در بین پژوهش‌های انجام شده در حوزه تشخیص گفتار ساختگی از نظر معیار ارزیابی EER قرار می‌گیرد.

کلمات کلیدی: تشخیص گفتار جعلی، یادگیری عمیق، تشخیص جعل عمیق صوتی.