Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection

Haowei Cheng, Candy Olivia Mawalim, Kai Li, Lijun Wang, and Masashi Unoki
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: {haowei.cheng, candylim, kai_li, lijun.wang, unoki}@jaist.ac.jp

***Abstract*—Deep-fake speech detection aims to develop effective techniques for identifying fake speech generated using advanced deep-learning methods. It can reduce the negative impact of malicious production or dissemination of fake speech in real-life scenarios. Although humans can relatively easy to distinguish between genuine and fake speech due to human auditory mechanisms, it is difficult for machines to distinguish them correctly. One major reason for this challenge is that machines struggle to effectively separate speech content from human vocal system information. Common features used in speech processing face difficulties in handling this issue, hindering the neural network from learning the discriminative differences between genuine and fake speech. To address this issue, we investigated spectro-temporal modulation representations in genuine and fake speech, which simulate the human auditory perception process. Next, the spectro-temporal modulation was fit to a light convolutional neural network bidirectional long short-term memory for classification. We conducted experiments on the benchmark datasets of the Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and the Audio Deep synthesis Detection Challenge 2023 (ADD2023), achieving an equal-error rate of 8.33% and 42.10%, respectively. The results showed that spectro-temporal modulation representations could distinguish the genuine and deep-fake speech and have adequate performance in both datasets.**

I. Introduction

The development of deep learning technology is rapidly advancing. It offers numerous conveniences to our daily lives, such as the production of audiobooks[1], the creation of intelligent speech robots [2], and even aiding those who have lost their voices due to throat disease or other medical conditions. Speech is considered the most essential and natural way for humans to convey both linguistic and non-linguistic information. However, the emergence of malicious fake speech generated by deep learning poses significant threats to societal stability and individual property security. This form of fake speech is commonly known as "deep-fake speech". It deceives both human listeners and automatic speaker verification systems. Furthermore, it carries the risk of spreading false and harmful information, including distorting politicians' statements [3]. Therefore, it is crucial to develop a reliable method for effectively detecting deep-fake speech.

Several challenges have been organized to advance the field of deep-fake speech detection. One of the most renowned global challenges is the automatic speaker verification and spoofing countermeasures challenge (ASVspoof) [4]. Its primary objective is to promote the development of robust countermeasures against such spoofing attempts. Another notable recent challenge is the audio deep synthesis detection (ADD) challenge [5]. It specifically focuses on detecting deep-fake audio in realistic scenarios. This challenge aims to address the unique challenges posed by deep-fake situations encountered in real-life settings.

Although many challenges and methods are proposed in deep-fake speech detection tasks, it is difficult for machines to accurately distinguish them. The main reason for this difficulty lies in the inherent difference between genuine speech and fake speech. Genuine speech not only contains speech information but also reflects human vocal system activity, including characteristics like glottal vibration. On the other hand, fake speech generated by machines lacks these human-like characteristics. This disparity makes it challenging for machines to accurately distinguish between the two types of speech. Common features struggle to effectively capture the unique patterns associated with genuine speech information and human vocal system activity, resulting in insufficient discriminative information for neural network training. Therefore, achieving effective detection of deep-fake speech can be accomplished by successfully separating these components.

To address this issue, we were inspired by the human auditory mechanism. The human auditory cortex possesses dynamic and adaptive properties that enable us to effectively differentiate between speech from humans and machines [6]. Building upon this understanding, a study [7] has revealed that neurons in the auditory cortex can decompose spectrograms into spectro-temporal modulation (STM) content. This finding has led to the development of the STM, which is a multi-scale representation for speech analysis and has been shown to explain various psychoacoustic phenomena [8]. Another study [9] proposed an STM-based method for audio classification inspired by human auditory mechanisms. By utilizing an auditory model to capture relevant features, these methods have demonstrated their effectiveness in classification. Therefore, it is reasonable to consider that STM representation has the potential to discriminate deep-fake speech and enhance detection accuracy.

This paper focuses on exploring the cues and effectiveness of the STM representation which is based on the human auditory mechanisms for detecting deep-fake speech. We conducted an investigation into the role of various feature expressions and

implemented STM representations. To enhance the ability of neural networks to capture feature information within STM representations. We combined STM with a light convolutional neural network and bidirectional long short-term memory (LCNN-BiLSTM), which is a model widely used in classification tasks. We performed experiments on two datasets: the Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and the Audio Deep synthesis Detection Challenge 2023 (ADD2023). The results clearly demonstrated that the STM representations were highly effective in distinguishing between genuine and fake speech, achieving adequate performance on both datasets.

II. Related work

A. ASVspoof challenge

To promote research and development in combating voice spoofing attacks, the ASVspoof challenge was first held in 2015 and has since become a significant platform in the field. The competition provides datasets containing various types of voice spoofing attacks, such as speech synthesis, playback, and voice conversion. The objective of the ASVspoof challenge is to evaluate and improve the robustness of ASV systems, enabling them to reliably distinguish between genuine and fake speech. Through the competition, researchers collaborate and propose innovative methods to address the challenges posed by voice spoofing attacks.

B. ADD challenge

However, there has been a recognition that many real-life scenarios have not been adequately covered in ASVspoof challenge. To address this gap, the Audio Deep synthesis Detection challenge (ADD) was motivated. The ADD challenge encompasses three tracks: low-quality fake audio detection, partially fake audio detection, and the fake audio game. These tracks aim to provide a more comprehensive evaluation of ASV systems' ability to detect various types of voice spoofing attacks, including both low-quality and partially manipulated audio samples. The inclusion of these tracks in the ADD challenge aims to advance research in the field and encourage the creation of more robust and efficient solutions to counter voice spoofing attacks.

C. Common methods and limitations

Many common methods have been proposed for detecting deep-fake speech, typically involving front-end feature extraction and back-end classification [10–16]. Baseline models in the two challenges applied linear frequency cepstral coefficients (LFCC) as feature and Gaussian mixture models (GMM) for classification [17]. Meanwhile, many features were explored for deep-fake speech detection, including Mel-frequency cepstral coefficients (MFCC) [18, 19], constant-Q cepstral coefficients (CQCC) [20, 21], Gammatone cepstral coefficients (GTCC) [22], etc. For classification, a variety of models are commonly employed, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), etc.

Despite many features are employed in deep-fake speech detection tasks, it is difficult for machines to accurately distinguish them. On the other hand, humans can distinguish genuine and fake speech through our sense of hearing. Taking inspiration from the human auditory mechanism, our approach involves exploring feature representations that not only capture the speech content but also the subtle cues associated with human vocal system activity.

III. Spectro-Temporal modulation representation

A. Definition

Temporal modulation refers to the changes in modulations over time in the spectrogram, while spectral modulation represents variations along the frequency axis. The concept of STM combines both temporal and spectral modulations simultaneously. In the field of auditory psychophysics and neuroscience, the auditory model is divided into two essential stages. One involves transforming the acoustic signal into an internal neural representation called an auditory spectrogram. Another analyzes this spectrogram to estimate the spectral and temporal modulation content using specialized filters that respond to specific modulations[23, 24]. This stage aims to separate different cues and characteristics associated with distinct auditory percepts [25–27]. It can be compared to the adaptive and masking properties of the human auditory system, where important information can still be perceived even in a noisy environment. Therefore, incorporating STM analysis provides a more comprehensive understanding of human perception. By examining STM, we can potentially uncover meaningful characteristics in the speech signal that can aid in the detection of deep-fake speech.

To obtain the STM representation from the original signal, a series of steps is followed. Initially, the input signal undergoes decomposition into frequency components using filterbanks. This process separates the speech signal into different frequency bands. Subsequently, squaring and low-pass filtering are used for computing power envelope from the output of the filterbank. In the final step, a two-dimensional spectro-temporal analysis is conducted on this power envelope to derive the STM spectrogram. The STM spectrogram thus provides a representation of the dynamic variations present in the speech signal across different spectral and temporal scales.

B. Investigation the role of feature expressions

In order to investigate the role of feature expressions, three filterbanks were employed to implement the STM independently. The Mel filterbank (Mel FB) and Gammatone filterbank (ERB FB) are both widely utilized filterbanks in the realm of speech signal processing [28]. Constant bandwidth filterbank (CBW FB) maintains a constant bandwidth across the frequency range.

1) Mel filterbank: The Mel FB is based on the Mel scale, which is derived from psychoacoustic experiments and provides a nonlinear mapping of frequency. Triangular-shaped filters are employed by the Mel FB, with each filter centered at a

specific Mel frequency and possessing a bandwidth determined by the adjacent filters. This triangular shape enables the Mel FB to approximate the perceptual frequency resolution of the human auditory system [29]. The formula to convert a linear frequency (f) to the Mel scale (m) is as follows:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right), \tag{1}$$

2) Gammatone filterbank: The ERB FB is designed to accurately model the response characteristics of the cochlea in the human auditory system [30, 31]. It utilizes filters based on the Gammatone function, derived from a combination of complex exponential functions and low-pass filters. These filters effectively capture the shape of cochlear filters and the frequency selectivity of the auditory system. As a result, the filterbank enhances the representation of low-frequency components with narrow bandwidths and reduces the presence of high-frequency components with wider bandwidths, as shown in Figure 1. The integration of the ERB scale further enhances the accuracy by approximating the frequency resolution of the human auditory system. This integration allows the ERB FB to better capture the spectral characteristics of auditory signals and align with human auditory perception [32].

In the ERB FB, the center frequencies are based on the specified upper and lower frequency limits and the number of channels. These center frequencies are proportional to the corresponding bandwidths of the filters [33]. The output obtained from the ERB FB is as follows:

$$g_k(t) = \mathrm{A}t^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k)t) \cos(2\pi f_k t), \tag{2}$$

where $At^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k t))$ is the amplitude term represented by the Gamma distribution, A, n, and b_f are the amplitude, filter order, and bandwidth of the filter respectively. We apply the fourth order Gammatone. The formula to convert a linear frequency (f) to the ERB scale is as follows:

$$\mathrm{ERB} = 24.7(4.37 f_k + 1), \tag{3}$$

where f_k is the k-th center frequency (in kHz) of filterbank.

3) Constant bandwidth filterbank: Unlike Mel FB and ERB FB, which have variable bandwidths. CBW FB has a fixed bandwidth for each filter, regardless of their center frequency. It is constructed by employing a consistent bandwidth parameter, whereby the center frequencies and channel count are computed using the provided lower and upper frequency limits.

To analyze the differences between genuine and fake speech signals as illustrated in spectrograms. Figure 2 provides visualizations of the spectrogram representations.

C. Procedure of STM analysis

In the front-end input, the speech signal $s(t)$ is first filtered by a bank of filters. The output of the k-th channel is given by

$$y_k(t) = g_k(t) * \mathrm{s}(t), \tag{4}$$

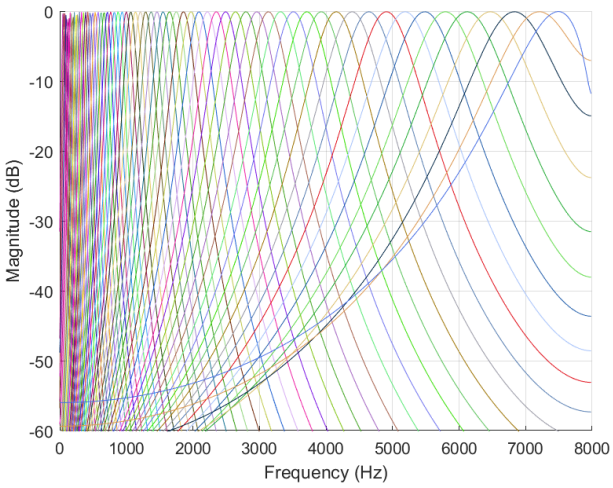

Fig. 1: Frequency response of ERB FB

where * represents the convolution, $g_k(t)$ is the impulse response of the k-th channel of filterbank. The power envelope is extracted by the Hilbert transform and squared. LPF represents a low-pass filter at cut-off frequency in 64Hz.

$$e_k^2(t) = \mathrm{LPF}\left[|\mathrm{y}_k(t) + j\mathrm{Hilbert}(y_k(t))|^2\right], \tag{5}$$

Finally, STM representation can be obtained by applying a two-dimensional Fourier transform to the squared envelope $e_k^2(t)$, as shown in Eq. (6). It is important to note that the result of the two-dimensional Fourier transform is typically a matrix comprising complex numbers, where each element consists of both real and imaginary parts. To obtain the STM representation utilized in this study, the absolute value of the result is taken. The STM representations of genuine and fake speech signals are shown in Figure 3.

$$\mathrm{STM} = \mathrm{2DFFT}(\log e_k^2(t)). \tag{6}$$

where 2DFFT represents a two-dimensional fast Fourier transform.

IV. Proposed Method

In this section, we present a deep-fake speech detection model that utilizes an LCNN and BiLSTM, as shown in Figure 4. LCNN is a convolutional neural network variant that is purposefully developed to strike a balance between computational complexity and performance. The advantage of LCNN is the implementation of a max feature map activation strategy, which involves using a max-out activation function. This characteristic enables faster training and inference times while minimizing the impact on overall performance.

In order to extract useful information from STM, deep learning models such as BiLSTM have been widely used. It can effectively model the temporal dependencies in the STM, which are critical for deep-fake speech detection tasks. Specifically, BiLSTM has a "memory" mechanism that allows it to keep track of past information and use it to inform the current prediction. During the task using BiLSTM, the speech

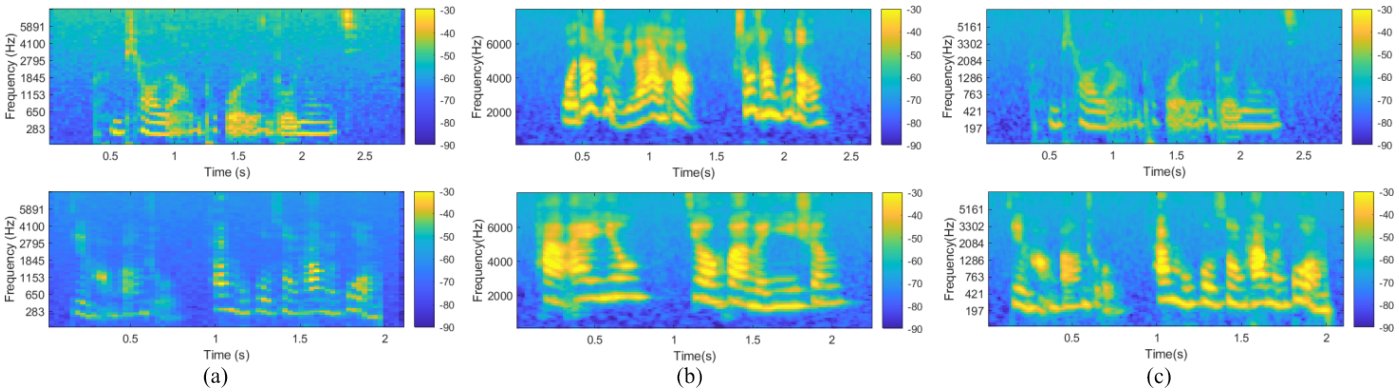

Fig. 2: Spectrograms of genuine (above) and fake (below) speech signals: (a) Mel FB, (b) CBW FB and (c) ERB FB.

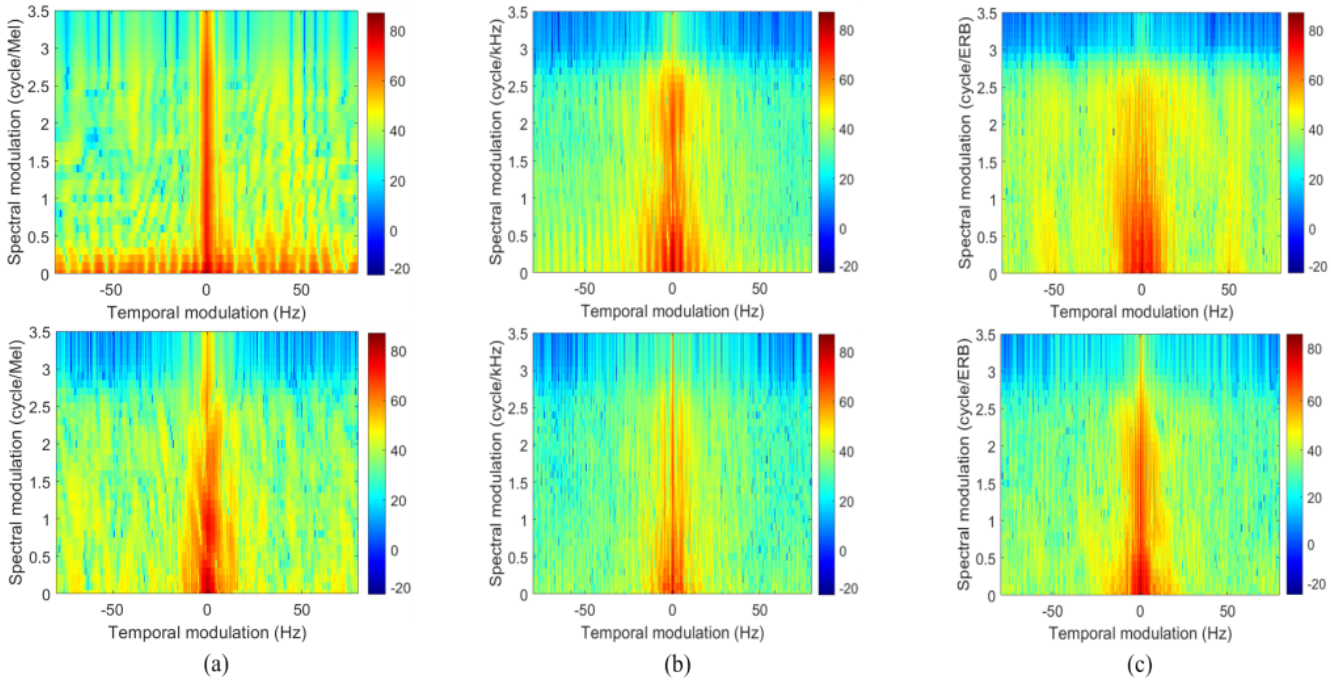

Fig. 3: STMs of genuine (above) and fake (below) speech signals: (a) Mel FB, (b) CBW FB, and (c) ERB FB.

feature sequences are individually fed into the hidden layers of both the forward LSTM (LSTM_F) and the backward LSTM (LSTM_B). This process generates two feature vectors that encapsulate the forward and backward information of the speech. Subsequently, the output vectors from these two layers are combined, forming a merged vector that is passed through two fully connected layers. Finally, the classification is performed by applying a sigmoid activation function to compute the score. This is particularly useful for distinguishing between genuine and fake speech, as speech signals often contain long-term dependencies. The dimensions of the BiLSTM layers are set to match the output dimensions of the LCNN. To optimize the model parameters, a binary cross entropy (BCE) objective function is utilized. The BCE objective function is defined as follows:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (7)$$

where N represents the total number of samples in the dataset, y_i and $\hat{y}_i$ denote the ground truth of the i-th training sample and its corresponding output probability from the model.

V. EXPERIMENT

A. Database and Metrics

In this study, two datasets were employed. The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof2019) pioneered the comprehensive treatment of all major attack types, including text-to-speech, voice conversion, and replay spoofing attacks, effectively covering real-world

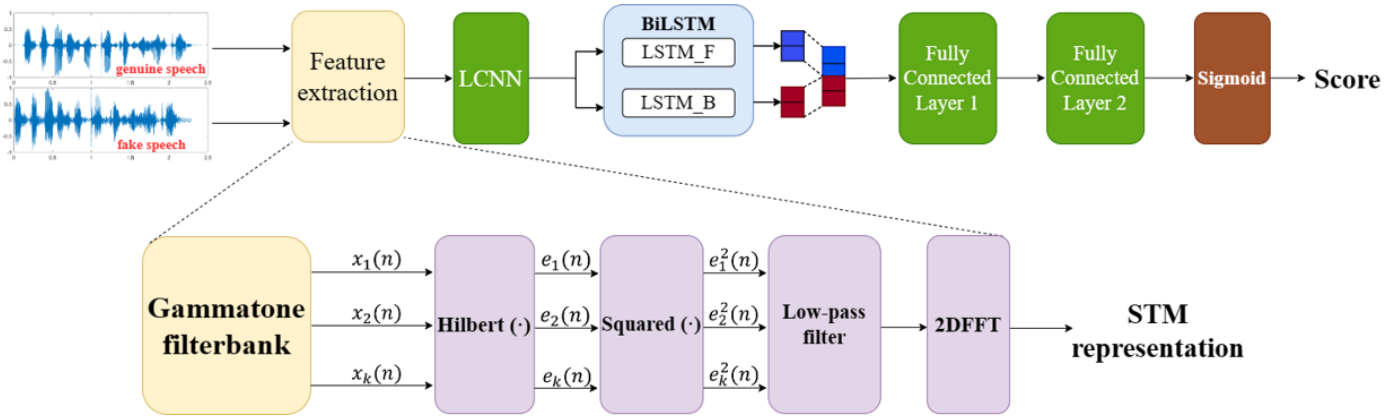

Fig. 4: Block diagram of the proposed method

TABLE I: statistic for datasets of the ADD2023. (Durations with three values denoted with minimum/average/maximum)

Dataset	Number of utterances			Duration (sec)
	Genuine	Fake	Total	
Training	3,012	24,072	27,084	0.86/3.15/60.01
Development	2,307	26,017	28,324	0.86/3.16/60.01
Evaluation	-	-	111,977	0.35/5.51/217.49

TABLE II: statistic for datasets of the ASVspoof2019. (Durations with three values denoted with minimum/average/maximum)

Dataset	Number of utterances			Duration (sec)
	Genuine	Fake	Total	
Training	2,580	24,072	26,625	0.65/3.42/13.19
Development	2,548	22,296	24,844	0.69/3.49/16.51
Evaluation	7,355	63,882	71,237	0.47/3.14/16.55

voice spoofing scenarios [34]. Another dataset used is the Audio Deep synthesis Detection challenge (ADD2023) [35].

The ADD2023 dataset (as shown in Table I) consists of Mandarin speeches with neutral emotions. The training and development sets have high signal-to-noise ratios (SNR), while the evaluation set has low SNR with various real-world background noises. The evaluation set lacks publicly accessible labels provided by the organizers, therefore the final scores are required to be submitted to the ADD2023 challenge's website CODALAB for online evaluation.

The ASVspoof2019 dataset is divided into three subsets: the training set, development set, and evaluation set, as presented in Table II. Notably, the evaluation dataset in ASVspoof2019 includes provided labels, facilitating local evaluation without the need to submit results for external assessment. The performance of the proposed method was evaluated using the equal error rate (EER). By comparing the results obtained from the two datasets, the objective was to prove the generalization ability and reliability of the proposed methods. This was achieved by utilizing a diverse and challenging collection of samples provided by the two datasets.

B. Experiment setup

The Mel FB, CBW FB, and ERB FB were implemented with consistent parameters. The frequency range was set from 50 Hz to 8000 Hz, utilizing 64 channels. This frequency range was chosen due to the typical perception of speech signals by the human auditory system. The lower limit of 50 Hz captures the fundamental frequency component of speech, while the upper limit of 8000 Hz includes high-frequency resonances and harmonics.

Setting the channel number to 64 aims to enhance spectral information and improve sound resolution. With 64 channels, a finer frequency division is achieved, enabling precise capture of speech characteristics across various frequency ranges. The increased channel count provides more frequency detail, leading to a more accurate representation of spectral features and capturing a wider range of speech features. However, it's important to consider the computational and memory costs associated with higher channel numbers, as they can impact real-time performance and computational efficiency. Thus, a careful balance was struck by selecting 64 channels. To accommodate the high-resolution STM representations, the TM domain underwent resampling at a rate of 1000 Hz, resulting in an STM representation size of [64, 1000].

Then the LCNN-BiLSTM model is trained using the labels, the batch size of all data is 64, and the epoch number is 30. An Adam optimizer with a learning rate of 0.0001 was used. Validation was performed using the development dataset, and the model achieving the lowest EER score was considered the best-performing one.

C. Comparison with different classic features

In addition to our proposed method, we conducted a comparative analysis by re-implementing three well-known features: MFCC, LFCC, and GTCC. This allowed us to assess the performance of our approach against these established methods

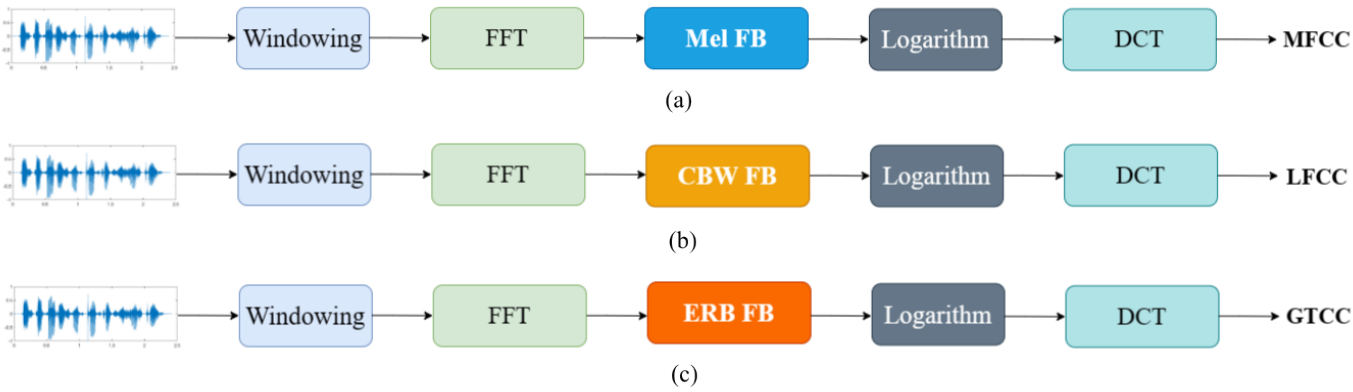

Fig. 5: Flow process diagram of classic features: (a) MFCC, (b) LFCC, and (c) GTCC.

[18–21].

Figure 5 illustrates the calculation diagram for the classic features. In the feature extraction stage, the input signal undergoes initial pre-processing, including windowing with a window length of 25 ms and a step length of 10 ms. The window type used is Hamming. Subsequently, a fast Fourier transform (FFT) is applied to the windowed signal with 512 points. This yields the Mel spectrum, Linear spectrum, and Gammatone spectrum after passing through the respective Mel FB, CBW FB, and ERB FB. Finally, the resulting spectrum is subjected to logarithmic transformation and discrete Cosine transform (DCT) to obtain the MFCC, LFCC, and GTCC. The back-end classifier used here is LCNN-BiLSTM, which shares the same architecture as the STM experiment.

VI. Results and Discussion

To identify effective features for deep-fake speech detection, we analyze the STM representation among genuine and fake speech. Subsequently, we applied STM and common features to LCNN-BiLSTM model and performed comparative experiments to evaluate the performance of the proposed method.

We conducted experiments using the ASVspoof2019 dataset (Table III) and the ADD2023 dataset (Table IV). The baseline model in ASVspoof2019 utilized LFCC and GMM, achieving an EER of 18.89%. In comparison, our STM-based approach with the ERB FB achieved an EER of 8.33%, representing a significant improvement of 10.56%. Notably, the STM (ERB FB) outperformed both STM (Mel FB) and STM (CBW FB) in terms of performance. These results emphasize the effectiveness of integrating STM and the advantages of using the ERB FB for deep-fake speech detection. The improved performance can be attributed to the STM's ability to capture fine-grained temporal and spectral details, facilitating more precise discrimination between genuine and fake speech samples.

In the evaluation of the ADD2023 dataset, we compared the performance of our proposed method with the baseline published by the organizer, which utilized LFCC-LCNN and achieved an EER of 70.37%. In our experiments, we first applied Mel FB, CBW FB, and ERB FB as input features to the classifier. The results showed EERs of 77.61%, 83.37%, and 73.34%, respectively. Then we conducted further comparison experiments and found that the classic features outperformed the filterbanks. Meanwhile, the STM representations performed better than the classic features. Specifically, STM based on ERB FB achieved an EER of 42.10%, outperforming not only the baseline but also common features like MFCC (53.36%) and GTCC (63.69%). These results indicate that STMs provide critical information to detect genuine and fake speech.

While our method has demonstrated successful deep-fake speech detection and achieved superior results, it is important to discuss its underlying principles and address the remaining issues. The STM can be considered as a cepstrum, which can capture the subtle cues of human vocal system activity information and speech information in a two-dimensional representation. In the case of genuine speech, the vocal system activity information tends to concentrate near the origin of the STM representation, while the speech information spreads around it. On the other hand, the STM representation of fake speech lacks this characteristic pattern observed in genuine speech. Natural speech produced by humans exhibits a more regular pattern in the STM representation due to the commonalities in the human vocal system. However, machine-generated speech lacks this consistent pattern across individuals, resulting in a less regular waveform in speech signals. Leveraging this feature, we can effectively distinguish between fake and genuine speech.

Therefore, we investigated the role of different feature expressions and implemented STMs. Then combined STM representations with an LCNN-BiLSTM model, and experiments on the datasets of the ASVspoof2019 and the ADD2023. The results showed that STM can effectively distinguish genuine and fake speech with good performance both in two datasets.

TABLE III: Comparative results using the ASVspoof2019 dataset

Methods	Equal Error Rate (%)	
	Development set	Evaluation set
STM (Mel FB)	0.04	9.79
STM (CBW FB)	0.09	13.46
STM (ERB FB)	**0.02**	**8.33**

TABLE IV: Comparative results using the ADD2023 dataset

Method	Equal Error Rate (%)	
	Development set	Evaluation set
Mel FB	0.26	77.61
CBW FB	0.31	83.37
ERB FB	0.23	73.34
MFCC	0.14	53.36
LFCC	0.19	66.52
GTCC	0.21	63.69
STM (Mel FB)	0.14	47.65
STM (CBW FB)	0.26	55.55
STM (ERB FB)	**0.09**	**42.10**

VII. CONCLUSION

This study showed the effectiveness of utilizing STM based on the ERB FB, inspired by the human auditory mechanism for deep-fake speech detection. First, we investigated the role of different feature expressions and implemented STM representations. We then developed a classifier by combining LCNN with BiLSTM and conducted experiments. Additionally, we compared the performance of three classical features with the STM method. The proposed method was evaluated on the ASVspoof2019 and ADD2023 datasets. Remarkably, our method exhibited significant results for ASVspoof2019 and ADD2023, achieving performance of 8.33% and 42.10% respectively. These results demonstrate the proposed method could effectively detect deep-fake speech. While the focus of this study was on evaluating the performance against baseline models, conducting comparisons with other existing deep-fake speech detection systems could provide a more comprehensive assessment of the proposed method's effectiveness. Future work will involve further investigation of the specific physics-based acoustic features that can be accurately captured and represented by the STM representation. Moreover, it is essential to expand the evaluation to encompass other state-of-the-art methods and to assess the system's performance across a diverse range of datasets.

ACKNOWLEDGEMENT

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233) and Grant-in-Aid for Transformative Research Areas (A) (23H04344).

REFERENCES

[1] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: An overview," in *Proc. Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, 2021, pp. 557–566.

[2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, 2010, pp. 537–541.

[3] R. Naika, "An overview of automatic speaker verification system," in *Proc. Intelligent Computing and Information and Communication, ICICC 2017*, pp. 603–610.

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

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 9216–9220.

[6] B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Adaptation of the human auditory cortex to changing background noise," *Nature communications*, vol. 10, no. 1, p. 2509, 2019.

[7] S. Shamma, "Encoding sound timbre in the auditory system," *IETE Journal of research*, vol. 49, no. 2-3, pp. 145–156, 2003.

[8] R. P. Carlyon and S. Shamma, "An account of monaural phase sensitivity," *Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 333–348, 2003.

[9] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–601.

[10] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting deepfake voice using explainable deep learning techniques," *Applied Sciences*, vol. 12, no. 8, p. 3926, 2022.

[11] P. A. Ziabary and H. Veisi, "A countermeasure based on cqt spectrogram for deepfake speech detection," in *Proc. International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2021, pp. 1–5.

[12] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9241–9245.

[13] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[14] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6359–6363.

[15] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[16] Z. Jiang, H. Zhu, L. Peng, W. Ding, and Y. Ren, "Self-supervised spoofing audio detection scheme." in *Proc. Interspeech*, 2020, pp. 4223–4227.

[17] Y. Ma, Z. Ren, and S. Xu, "Rw-resnet: A novel speech anti-spoofing model using raw waveform," *arXiv preprint arXiv:2108.05684*, 2021.

[18] R. K. Bhukya and A. Raj, "Automatic speaker verification spoof detection and countermeasures using gaussian mixture model," in *Proc. IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2022, pp. 1–6.

[19] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning." in *Proc. Interspeech*, 2019, pp. 1048–1052.

[20] M. Sahidullah, H. Delgado, M. Todisco, A. Nautsch, X. Wang, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pp. 339–385, 2023.

[21] S. Ravindran and K. Geetha, "An overview of spoof detection in asv systems," *ECS Transactions*, vol. 107, no. 1, p. 1963, 2022.

[22] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, "Voice spoofing detector: A unified anti-spoofing framework," *Expert Systems with Applications*, vol. 198, p. 116770, 2022.

[23] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE transactions on speech and audio processing*, vol. 3, no. 5, pp. 382–395, 1995.

[24] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[25] F. Samson, L. Mottron, B. Jemel, P. Belin, and V. Ciocca, "Can spectro-temporal complexity explain the autistic pattern of performance on auditory tasks?" *Journal of autism and developmental disorders*, vol. 36, pp. 65–76, 2006.

[26] S. M. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature neuroscience*, vol. 8, no. 10, pp. 1371–1379, 2005.

[27] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[28] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[29] R. S. Chavan and G. S. Sable, "An overview of speech recognition using hmm," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 6, pp. 233–238, 2013.

[30] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Replay attack detection in automatic speaker verification using gammatone cepstral coefficients and resnet-based model," *Journal of Signal Processing*, vol. 26, no. 6, pp. 171–175, 2022.

[31] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Proc. IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[32] T. Irino and M. Unoki, "A time-varying, analysis/synthesis auditory filterbank using the gammachirp," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, 1998, pp. 3653–3656.

[33] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[34] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[35] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.