Study on Deep-fake Speech Detection Based on Spectro-Temporal Modulation Representation

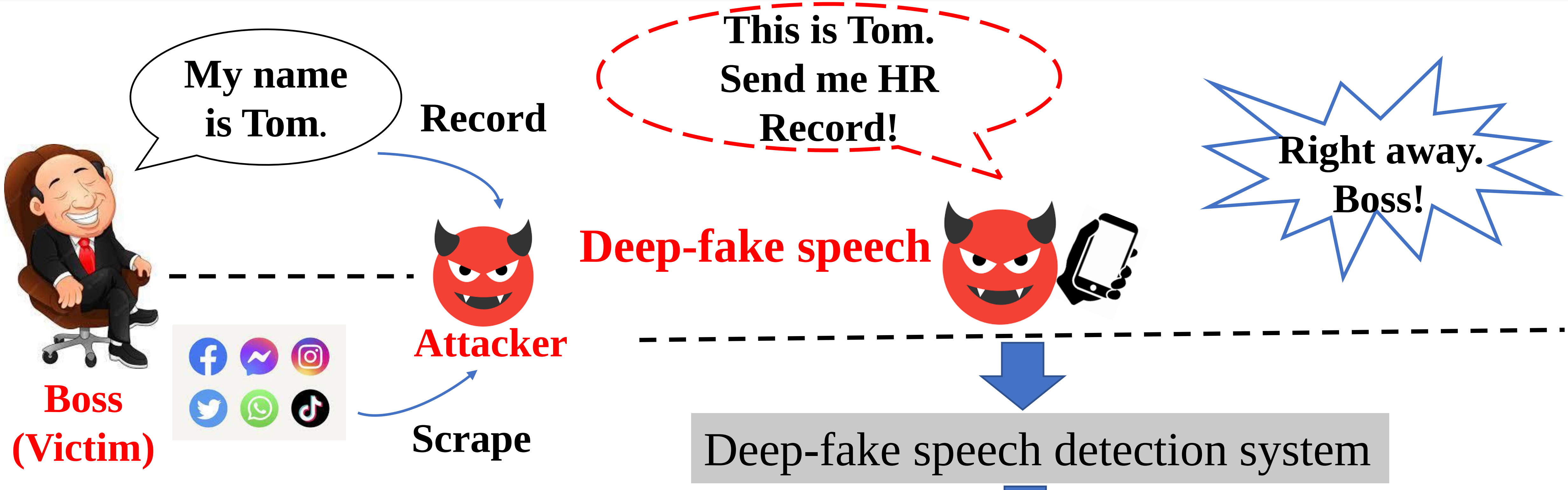
Haowei Cheng, Candy Olivia Mawalim,
Kai Li, Masashi Unoki (JAIST)

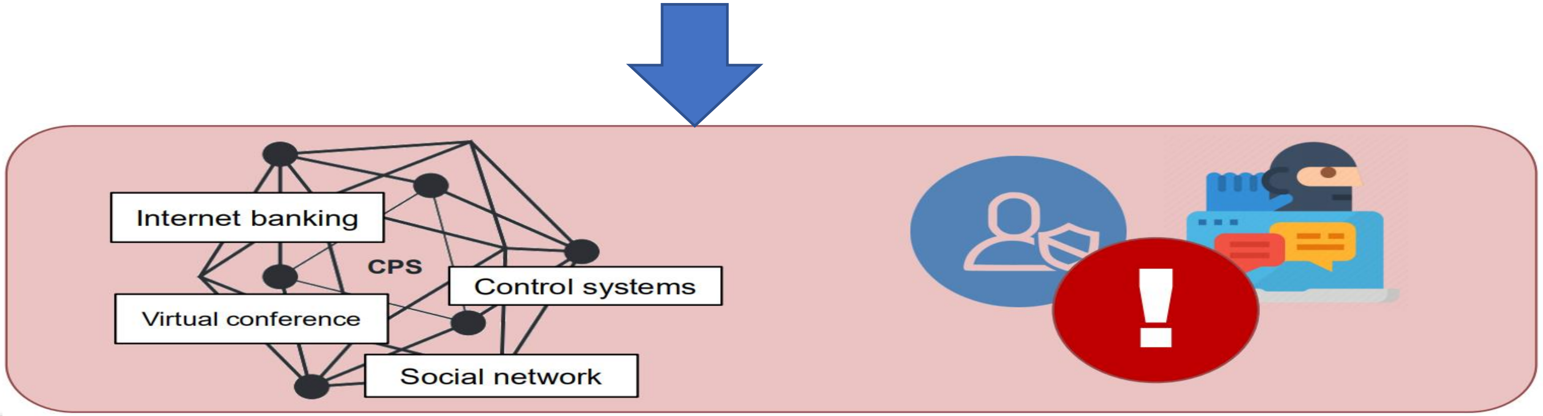
2023/09/02

JHES2023
2023年度電気・情報関係学会北陸支部連合大会

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

Research background





Attacker steals Boss's speech via recording or social media.

Research background

Spectro-temporal Modulation (STM) combines spectral and temporal modulations, providing a way to mimic the dynamic characteristics of the human auditory system. STM-based features can better represent the perceptual aspects of speech signals.

Prof. Shamma revealed that neurons in the auditory cortex system can decompose spectrograms into STM representations. This finding has been shown to explain various psychoacoustic phenomena [1].

Dr. Carlyon introduced an STM-based method for audio classification, and the approach has demonstrated its effectiveness [2].

[1] 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.
[2] 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, 2

Significance

It is helpful for/to:

- Reducing the negative impact of malicious production or dissemination of deep-fake speech in real-life scenarios.
- Provide theoretical support for a deep-fake speech detection system.

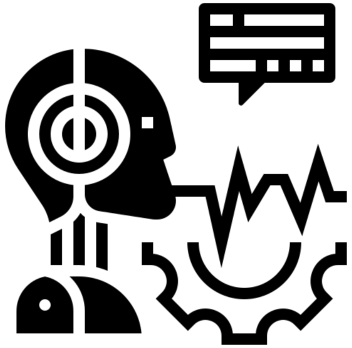
Research issue

Although many Challenges and methods are proposed in deep-fake speech detection tasks, it is difficult for machines to precisely distinguish them [3].

Speech content

Human vocal system activity

Fake speech generated by machines lacks these human-like characteristics.



[3] 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

Purpose

Developing an effective technique for detecting deep-fake speech, by analyzing how the human auditory mechanism perceives and processes speech.

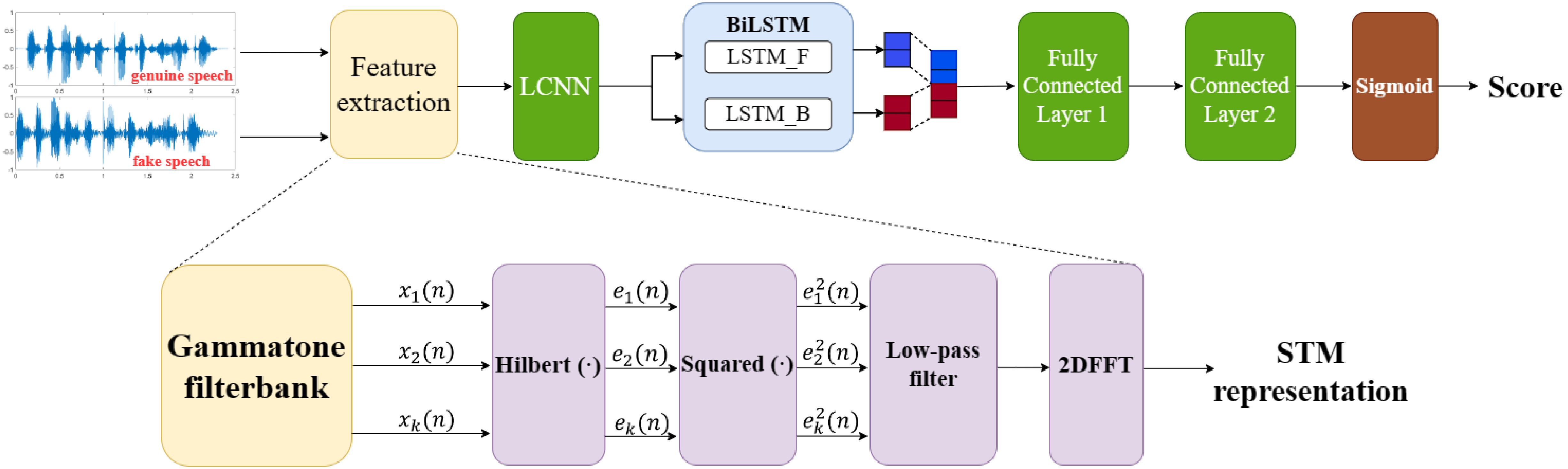
The ultimate goal is to mitigate the negative impact of maliciously produced or disseminated fake speech in various real-life scenarios.

Methodology

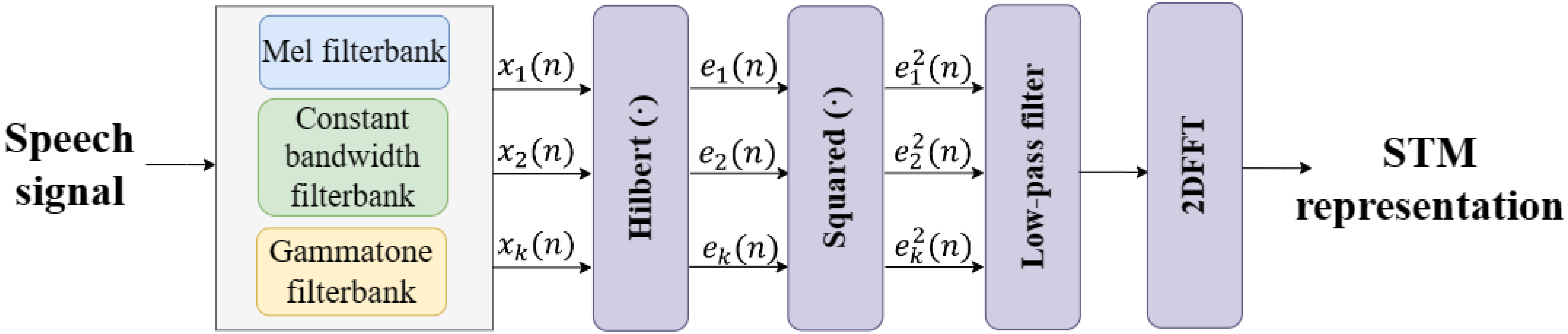
- **Stage 1** Investigate useful features and cues of speech perception
- **Stage 2** Construct prototype of the proposed method for detecting deep-fake speech
- **Stage 3** Evaluate the proposed method and conduct comparison experiments compare with other methods
- **Stage 4** Solve the remaining problems and refine the proposed method

Proposed method

Proposed method including two parts: **Spectro-temporal Modulation (STM) extraction** and **Identification**.

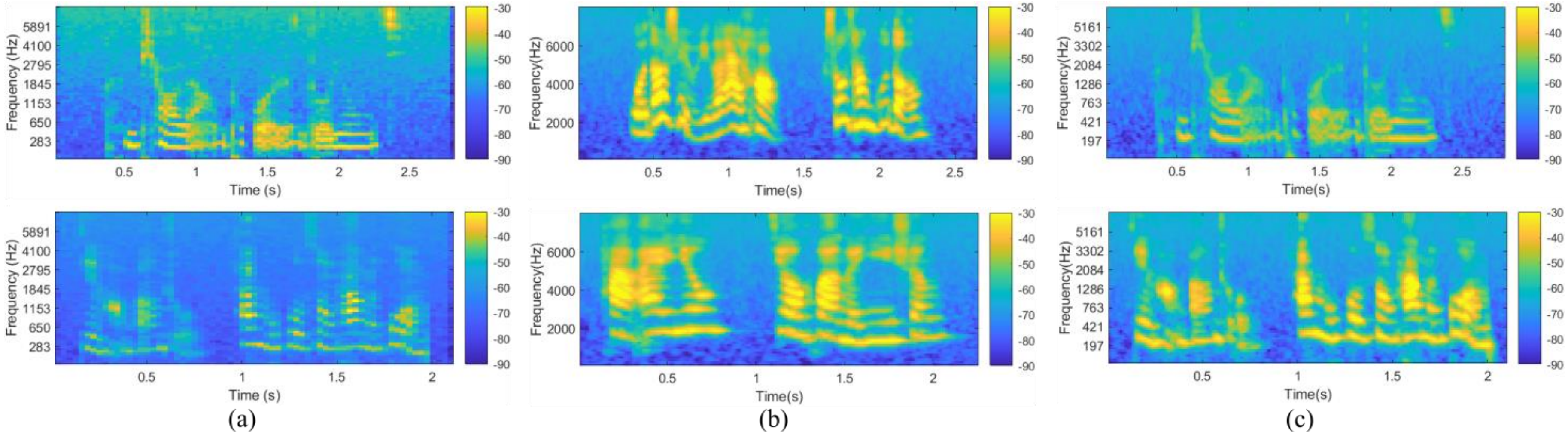


Investigation of feature expressions



Genuine speech

Fake speech

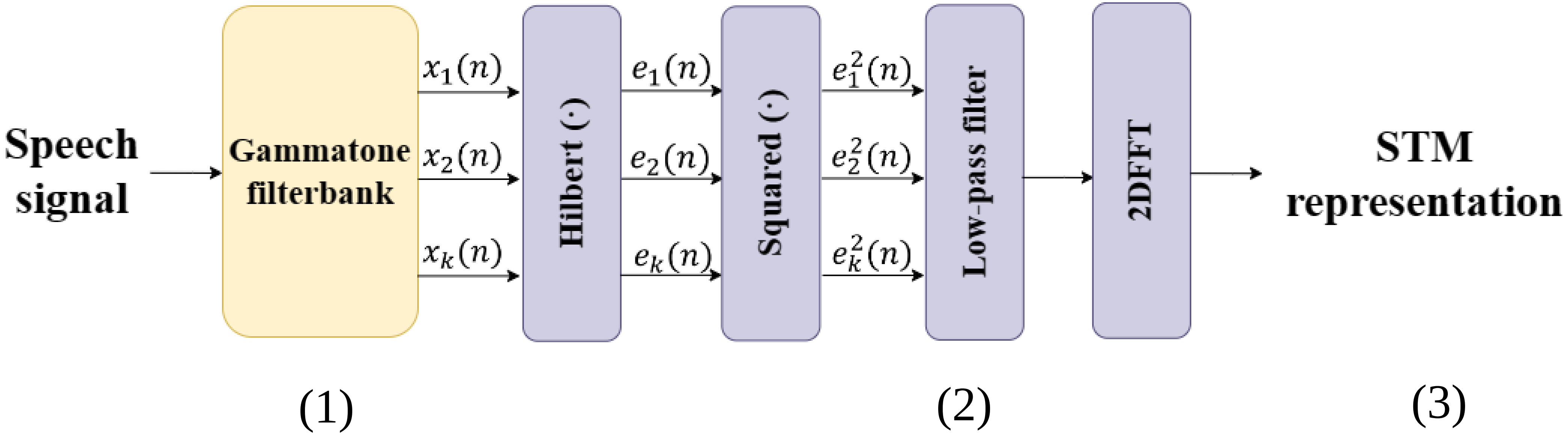


Mel Filterbank (**Mel FB**)

Constant Bandwidth Filterbank (**CBW FB**)

Gammatone Filterbank (**ERB FB**)

Procedure of STM extraction



(1) $y_k(t) = g_k(t) * s(t)$

(2) $e_k^2(t) = \mathrm{LPF}\left[\left| y_k(t) + j\mathrm{Hilbert}\big(y_k(t)\big)\right|^2\right]$

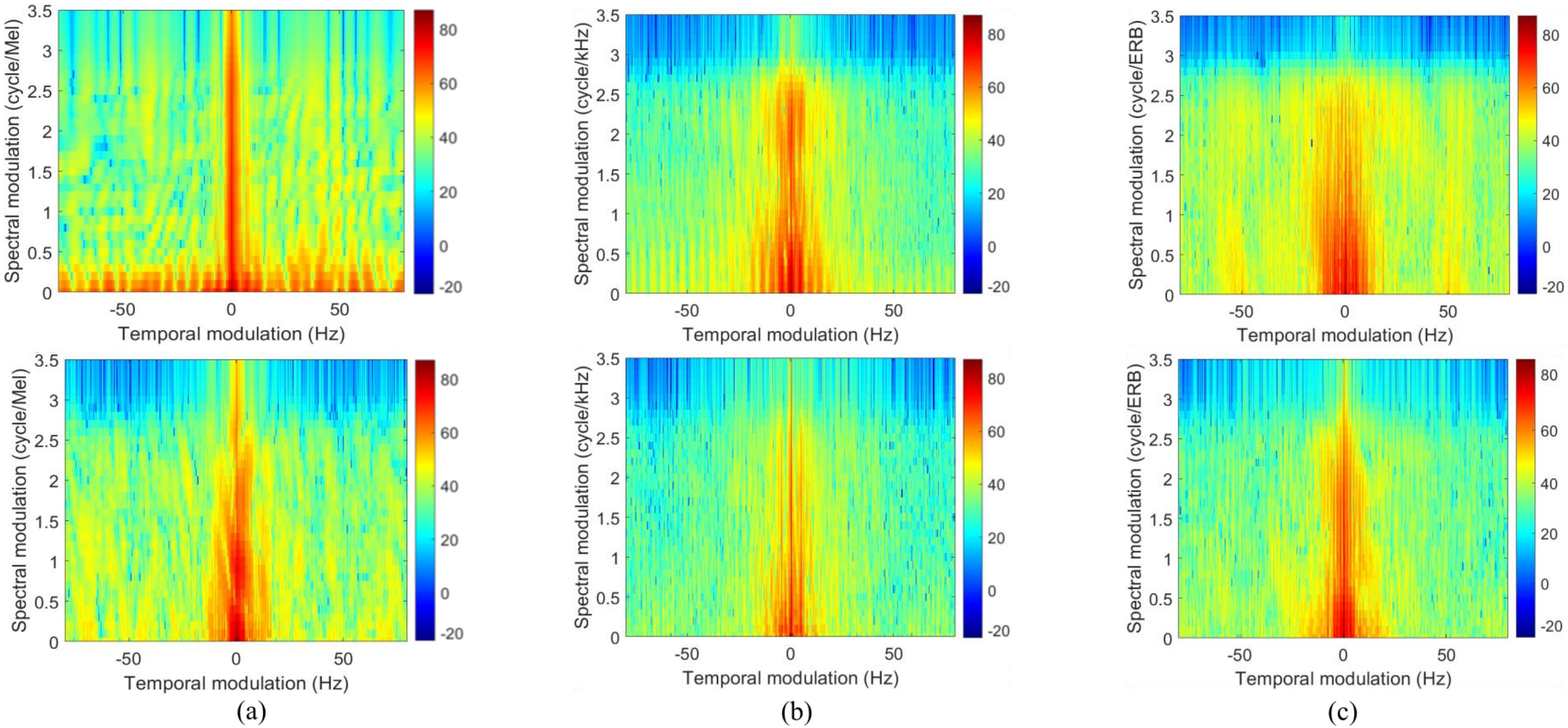
(3) $\mathrm{STM} = 2DFFT\left(\log e_k^2\,(t)\right)$

* : convolution
LPF: low-pass filter
Hilbert: Hilbert transform
2DFFT: 2-dimentional fast Fourier transform

STM representations

Genuine speech

Fake speech

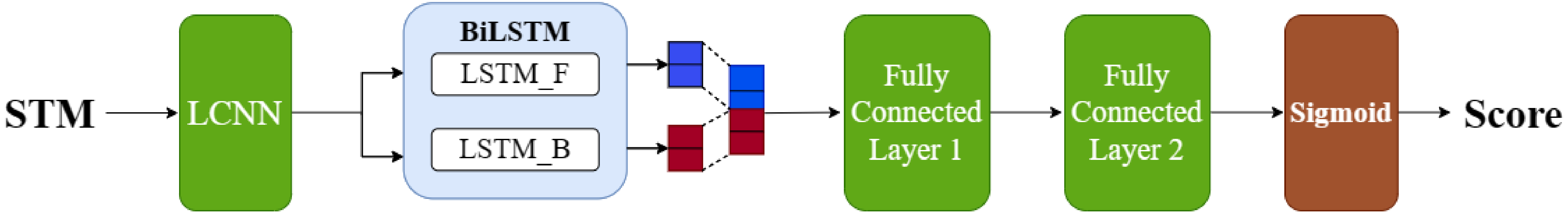


Mel FB

CBW FB

ERB FB

Identification



Batch size: 64
Epoch number: 30
Adam optimizer (learning rate): 0.0001
Loss function: Binary cross entropy

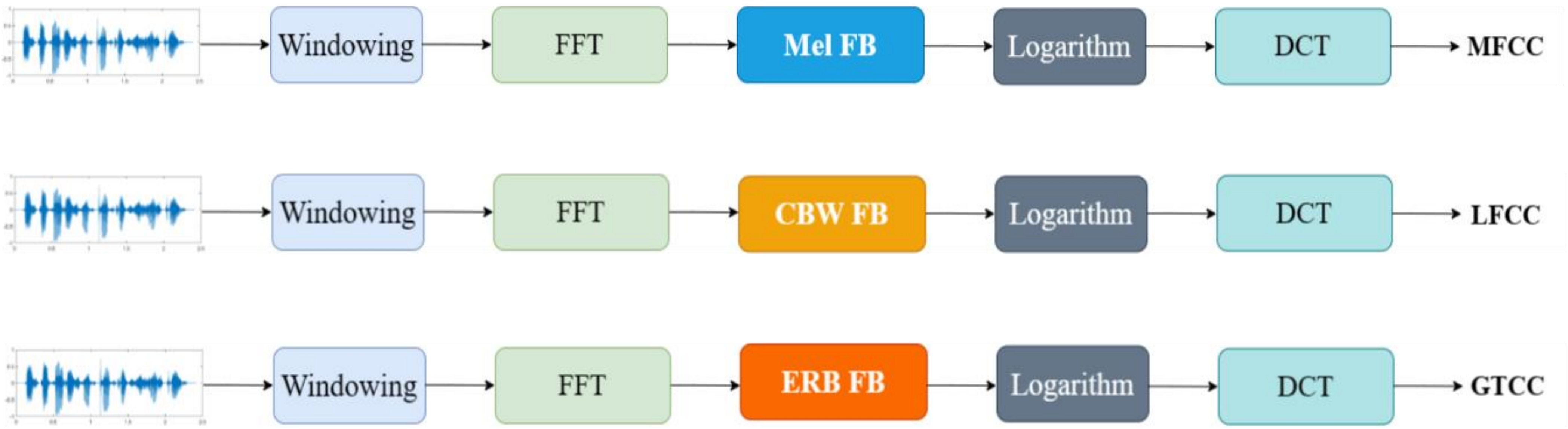
LCNN: light convolutional neural network
BiLSTM: bidirectional long short-term memory
LSTM_F: forward LSTM
LSTM_B: backward LSTM
BCE: binary cross entropy

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Comparison experiments

We conducted comparative experiments using three well-known features: MFCC, LFCC, and GTCC.

The comparison experiments allow us to assess the performance of our approach against these methods.



Datasets and metrics

Datasets:

1. The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof)

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

2. Audio Deep synthesis Detection challenge (ADD)

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

Metric:

Equal Error Rate (EER) is a performance metric for binary classification tasks. **The smaller value of EER has the better performance.**

Experiments results

Comparative results using the ASVspoof2019 dataset (above) and ADD2023 (below)

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**

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**

- In different feature expressions, the result of ERB FB is better than Mel FB and CBW FB.
- STM representation based on ERB FB shows the better results than other approaches (MFCC, LFCC, GTCC).
- The results indicate that STMs could effectively distinguish between genuine and fake speech.

Conclusion

- ◆By analyzing the concept of STM representation, we gain valuable insights into how the human auditory mechanism perceives and processes speech.
- ◆We introduced a LCNN-BiLSTM model that utilizes STM representations for efficient deep-fake speech detection. The approach demonstrated better performance compared to common features.
- ◆Our work offers theoretical support for a fake speech detection system, which has the potential to reduce the negative impact of maliciously produced or disseminated deep-fake speech in real-life scenarios.