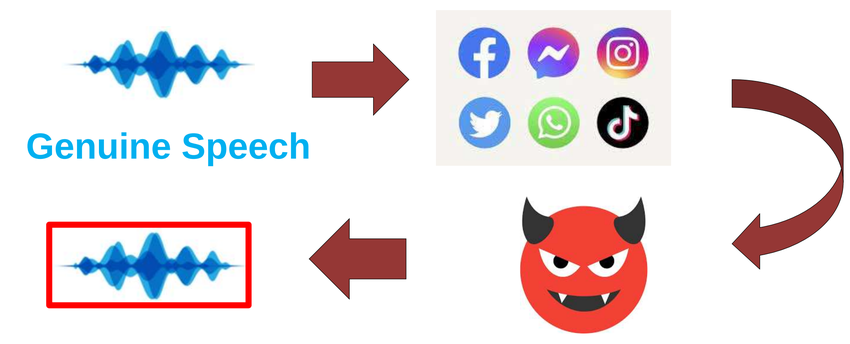
Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection

P4-02

Haowei Cheng, Candy Olivia Mawalim, Kai Li, Lijun Wang, Masashi Unoki
Japan Advanced Institute of Science and Technology, Japan

BACKGROUND



Spectro-temporal Modulation (STM) combines spectral and temporal modulations, providing a way to mimic the dynamic characteristics of the human auditory system.

(Shamma, 2004) revealed that neurons in the auditory cortex system can decompose spectrograms into STM representations. This finding has been shown to explain various psychoacoustic phenomena.
(Carlyon, 2005) introduced an STM-based method for audio classification, and this approach has demonstrated its effectiveness.

ISSUES

- Speech content
- Human vocal system activity

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

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

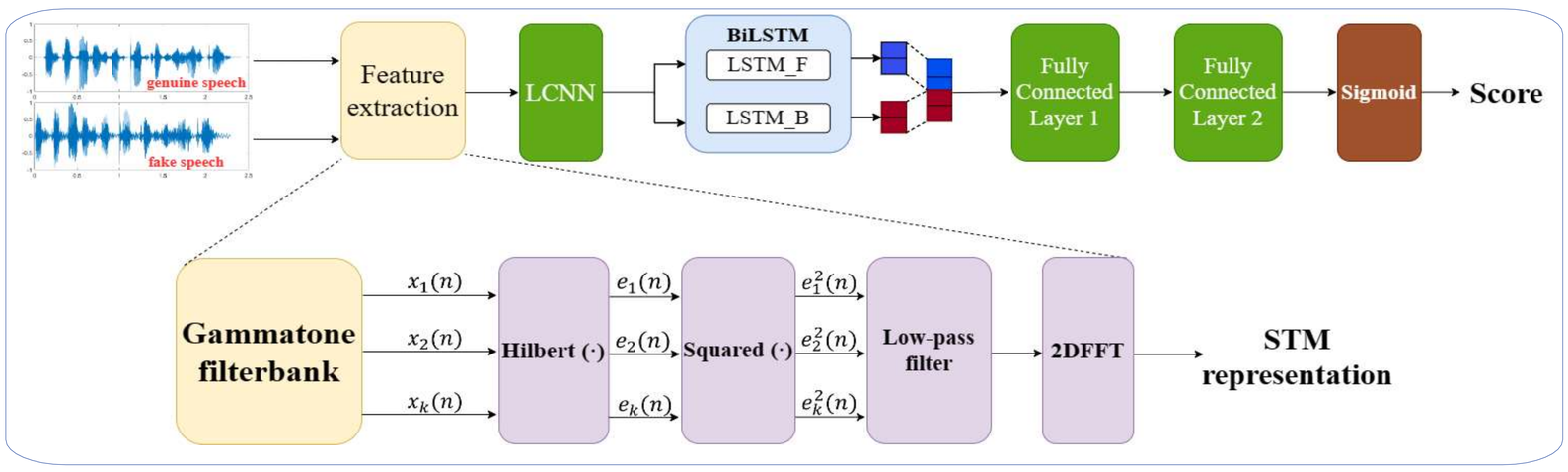
DATASETS & METRIC

1. The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof)
2. Audio Deep synthesis Detection challenge (ADD)

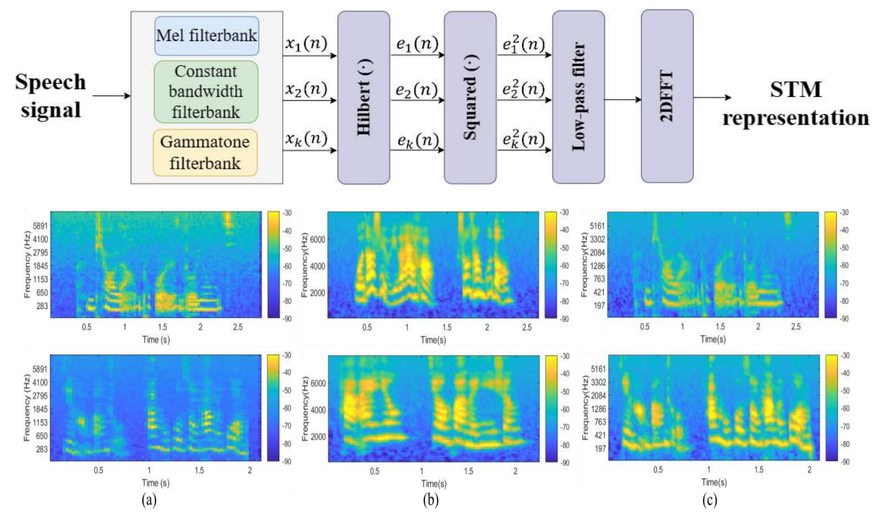
Equal Error Rate (EER)

- It is a performance metric when False Accept Rate (FAR) and False Rejection Rate (FRR) are equal.
- The smaller value of EER has the better performance.

PROPOSED METHOD

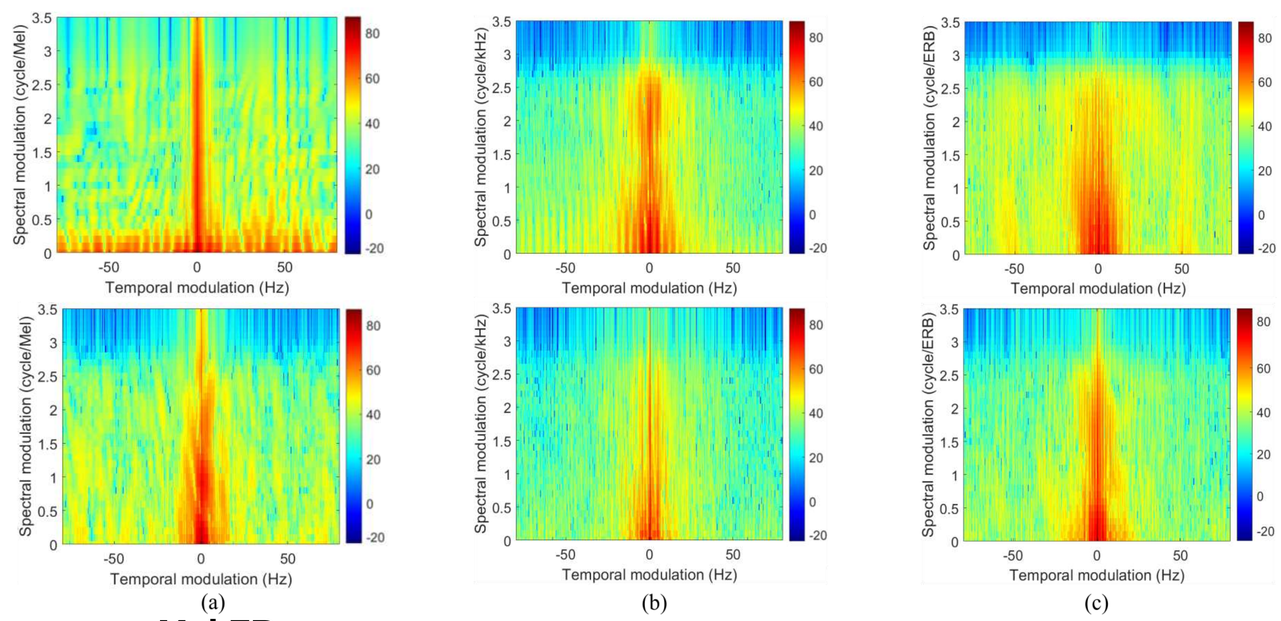


Investigation of feature expressions

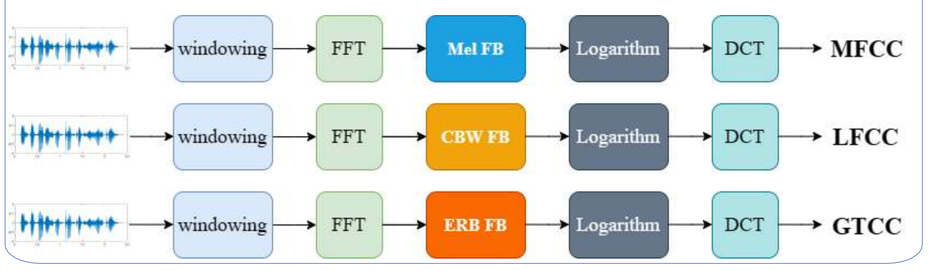


Mel FB CBW FB ERB FB

STM REPRESENTATION



Comparison experiments



EVALUATION

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.