2023年度電気・情報関係学会北陸支部連合大会

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

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

1. Background

Deep-fake speech refers to the manipulation or generation of human-like speech content using deep learning techniques. The content appears to be spoken by a particular individual, even though they did not say those words. It carries the risk of spreading false and harmful information, including distorting politicians' statements. Although humans could relatively be easy to distinguish between genuine and fake speech due to human auditory mechanisms, it is difficult for machines to distinguish them correctly.

To address this issue, we investigated spectro-temporal modulation (STM) representations, which simulate the human auditory perception process, it can capture the difference between human speech and machine-generated speech. This paper focuses on exploring the cues and effectiveness of STM for detecting deep-fake speech.

2. Proposed method

STM representation: Temporal modulation refers to changes in modulations over time in the spectrogram, while spectral modulation represents variations along the frequency axis. STM combines both temporal and spectral modulations, providing a comprehensive representation of the dynamic characteristics of a speech signal.

Gammatone filterbank: The Gammatone filterbank (ERB FB) is designed to model response characteristics of the cochlea in the human auditory system [1]. It enhances the representation of low-frequency components with narrow bandwidths and reduces the presence of high-frequency components with wider bandwidths. This characteristic allows ERB FB to better capture the spectral characteristics of speech signals and align with human auditory perception [2]. The output obtained from the ERB FB is as follows:

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

where $\mathrm{A}t^{(n-1)} \exp\big(-2\pi b_f \mathrm{ERB}(f_k)t\big)$ is the amplitude term represented by the Gamma distribution, A, n, and $\mathrm{b}_f$are the amplitude, filter order, and bandwidth of the filter respectively. The formula to convert a linear frequency (f) to the ERB scale is given by $\mathrm{ERB} = 24.7(4.37f_k + 1)$, where f_k is the $k-th$ center frequency (in kHz) of filterbank.

STM extraction process: As shown in Figure 1, the speech signal $s(t)$ is filtered by ERB FB. The output of the $k - th$ channel is given by $y_k(t) = g_k(t) * s(t)$. Then the power envelope is extracted by the Hilbert transform and squared. LPF represents a low-pass filter at a cut-off frequency of 64Hz.

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

Finally, STM representation can be obtained by applying a two-dimensional Fourier transform to the squared envelope $e_k^2(t)$.

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

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

Identification: To extract useful information from STM, the STM was fit to a light convolutional neural network bidirectional long short-term memory for classification.

Table 1 Comparative results using 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 2 Comparative results using ADD2023 dataset.

Methods	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
STM (Mel FB)	0.14	47.65
STM (CBW FB)	0.26	55.55
STM (ERB FB)	**0.09**	**42.10**

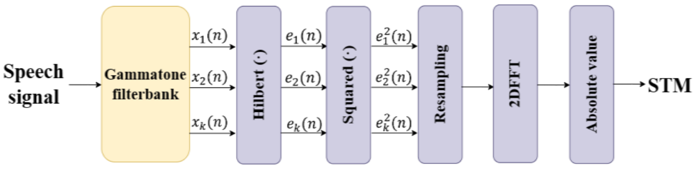

Figure 1 Block diagram of STM extraction

3. Evaluation

The experiments were conducted on the benchmark datasets of Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and Audio Deep synthesis Detection Challenge 2023 (ADD2023). The role of feature expressions including Mel filterbank (Mel FB), Constant bandwidth filterbank (CBW FB), and ERB FB was investigated. STM representations were implemented based on these feature expressions. The results, as presented in Table 1 and Table 2, showed the proposed method achieved an equal-error rate of 8.33% and 42.10%.

4. Conclusion

The results demonstrated that the STM based on the human auditory mechanism was capable of distinguishing between genuine and deep-fake speech. These representations exhibited satisfactory performance in both datasets, providing critical information for the detection of genuine and fake speech.

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).

Reference

[1] Patterson, R.D., Nimmo-Smith, I., Holdsworth, J. and Rice, P., "December. An efficient auditory filterbank based on the gammatone function," *Proc.* IOC Speech Group on Auditory Modelling at RSRE, 2(7), 1987.

[2] Moore, B.C. and Glasberg, B.R., "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am., 74(3), 750-753, 1983.