Deepfake algorithm recognition through multi-model fusion based on manifold measure

Ye Tian[1,*], Yunkun Chen[1], Yuezhong Tang[1] and Boyang Fu[1]

[1]The 3rd Research Institute of China Electronics Technology Group Corporation, No.B7 Jiuxianqiao North Road, Chaoyang District, Beijing, China

Abstract

This paper describes a deepfake algorithm recognition system submitted to the Audio Deep Synthesis Detection (ADD) Challenge Track 3, which aiming to recognize the algorithms of the deepfake utterances. Given the complex noise present in the testing data and the existence of unknown deepfake algorithms, we propose a manifold-based multi-model fusion approach for open-set recognition. This approach constructs a manifold space to fuse the deep embedding features extracted by different models and computes the geodesic distance between the manifold spaces of different deepfake algorithms to distinguish unknown deepfake methods. Experimental results demonstrate the effectiveness of the proposed strategy in multi-model fusion. The proposed system obtained the F1-score of 0.7934 in ADD Track 3 testing.

Keywords

Deepfake algorithm recognition, model fusion, manifold space

1. Introduction

Currently, the naturalness and similarity of synthetic speech are continuously improving, and in some conditions, they are comparable to those of human speech [1, 2]. While speech generation technology provides convenience for intelligent applications, it also brings threats to information cognition and social security. To safely cope with generative audio, fake audio detection has become one of the hot research spots [3]. Meanwhile, it is necessary to trace the source of fake audio.

In the field of audio genuine/fake recognition, feature and classifier design have always been hot topics. From the perspective of scheme structure, they can be mainly divided into three types of architectures: handcrafted features with classifiers, end-to-end classifiers, and pre-trained feature extractors with classifiers. Handcrafted features such as constant-Q cepstral coefficients (CQCC) and linear frequency cepstral coefficients (LFCC) are frequently employed in this field [4]. Additionally, residual networks [5] and LCNN [6] are common classifiers utilized for this purpose. End-to-end networks, such as rawnet2 [7], and pre-trained models, such as Wav2Vec 2.0 and WavLM Large [8], are also employed in this area of research. Different models extract information from different perspectives, and model fusion is a way to improve overall performance. However, there is not only complementary information between different models, but also information redundancy and interference, and an improper fusion strategy can instead degrade the overall performance [9].

Given the rapidly changing means of speech generation driven by market demand, it is difficult to include all generation means during the training phase of traceability models, which leading to an open-set recognition problem. Literature [10] provides an overview of current open-set recognition methods. Overall, these methods have their own characteristics, and their effectiveness needs to be comprehensively evaluated based on the actual application scenarios.

To address these problems, considering the specific needs in ADD 2023 Track 3, we propose a multi-model fusion method. Inspired by OpenMax [11] and manifold space used in recognition tasks [12], we regard the inputs of the classifier layer as the extracted discriminative features, and achieve fusion by constructing the manifold spaces of different labels and calculating the geodesic distances between the manifold spaces.

The main contributions of this study can be summarized as follows:

(1) We propose a manifold-based multi-model fusion approach. It achieved 0.7352 in F1-score, ranking 5th on track 3 in ADD 2023 during competition, and so far it reached 0.7934 in F1-score, ranking 3rd on the competition results list.

(2) We explore three strategies for model fusion, and demonstrate the effectiveness of manifold-based feature-level fusion and score-level fusion by inference augmentation.

(3) We describe and discuss the proposed method as well as the problems encountered in the competition and the ideas to solve them.

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R

[*]Corresponding author.
tianye_cetc3@163.com (Y. Tian); yunkun.chen@connect.polyu.hk (Y. Chen); tangyuezhong@cetc.com.cn (Y. Tang); y18311368068@163.com (B. Fu)

The rest of this paper is organized as follows: Section 2 describes the task. Section 3 presents the related work and illustrates our proposed method. Results and discussions are reported in Section 4. Finally, the paper is concluded in Section 5.

2. Task description and data

The Audio Deep Synthesis Detection (ADD) Challenge Track 3 [13] aims to recognize the algorithms of the deepfake utterances. The testing dataset includes known and unknown algorithms of the fake ones. The training and developing sets include 7 classes (1 real and 6 counterfeit), the 7 categories are labeled 0, 1, 2, 3, 4, 5, 6. The testing set includes 8 classes (the 7 classes included in the training and developing sets + 1 unknown counterfeit).

There are 22,400 training data, 8,400 developing data, and 79,490 testing data. In addition to containing unknown categories, the noise of the testing data is much more complex than the training data. It is clear that this challenge is focused on improving the generalization ability of the model based on limited training data.

Metrics for this track is the macro-average precision, recall, and F1-score.

3. System description

In order to improve the performance of the system on the testing set, some measures have been taken in terms of the data layer, feature layer, model layer and finally the score calculation, which are described in detail below.

3.1. Data augmentation

First, by observing the training data, we found that the audio were sampled at 16 or 24 kHz, and the volume of the audio varies relatively widely. Thus, the whole audio were uniformly resampled to 16 kHz and normalized.

Then, by examining the testing data, compared to the relatively clean training and developing data, the noise interference in the test data was more complicated, then data augmentation was performed on training and developing data with MUSAN [14] dataset. And the SNR was set randomly among 15 30dB.

Finally, some completely silent segments with zero-volume were found in these datasets. Although this may be a characteristic of some deepfake methods, the silent segments that appear at the beginning and the end of the audio were cropped out considering the generalized application of the model.

3.2. Features

To handle the complexity of the testing data, we explored three categories of features: raw waveform, hand-crafted features, and pre-trained features. Our expectation was that a combination of these features would be able to capture the divergences among different deepfake algorithms.

Based on the findings in literature [8], it has been demonstrated that anti-spoofing systems can achieve good performance by using raw waveforms with an end-to-end network architecture. In our work, a unified audio duration of 3s was applied in subsequent processing with truncation or padding.

Hand-crafted features are extracted based on specific knowledge, in contrast to raw waveforms. Several features are widely used in anti-spoofing, such as constant-Q cepstral coefficients (CQCC), linear frequency cepstral coefficients (LFCC), and log power magnitude spectrogram (Spec) [4]. While these features have demonstrated utility in anti-spoofing, we chosed to use LFCC as the hand-crafted feature in track 3 based on our previous tests with the ASVSpoof2019 dataset.

Due to the complexity of the testing data and the scarcity of available training data, we utilized a pre-trained model to extract essential speech features. Recently, some pre-trained speech models, including Wav2vec 2.0 [15], HuBERT [16] and WavLM [17], have demonstrated significant performance improvements in downstream tasks such as Automatic Speech Recognition, Text-to-speech and Voice Conversation. As some experiments have shown that HuBERT performs comparably or better than the current leading Wav2vec 2.0 on various benchmarks, we utilized a HuBERT model as a feature extractor and fed raw waveform as input to the model.

3.3. Deep recognition network

In our work, we utilized three different deep networks: rawnet2, SE-Res2Net50, and HuBERT.

rawnet2 [18] is an end-to-end network that is trained on raw audio and consists of one sinc layer, six residual blocks with attention mechanism, gate recurrent units (GRU), and two fully-connected layers. In our work, a softmax function was added to the output layer to produce seven-class predictions corresponding to the categories in the training dataset. The model was trained for 100 epochs with a batch size of 32 and a learning rate of 0.0001.

SE-Res2Net50 [19] is an improved version of the ResNet [20] model that combines squeeze-and-excitation (SE) with Res2Block. We trained the model using LFCC features with cross-entropy as the loss function and Adam as the optimizer with default parameters. The model was trained for 40 epochs with a batch size of 48 and a learning rate of 0.0002.

HuBERT is a self-supervised learning pre-trained model and is available in several versions. We utilized the chinese-hubert-large [21] model, which was trained using the WenetSpeech train L subset. Following the final layer of the model, we added two fully-connected layers and a softmax function to generate predictions. To mitigate the limitation of computing resources, we trained the model with a batch size of 24 for 40 epochs.

To ensure the best performance, we selected the final model for testing from the above mentioned models with the highest F1-score in the developing dataset.

3.4. Manifold space and distance

To classify the categories of deepfake audio and identify unknown deepfake means, we adopted the manifold space and manifold distance. Firstly, the manifold space of each deepfake category was constructed using the ONPE method [22]. Then, the spatial geodesic distance [23] between different manifold spaces was calculated using equation (1) and inverted to serve as a similarity indicator. Finally, the softmax value was calculated using equation (2)-(4) as the final decision score.

$$d(S_1, S_2) = \|\Theta\|_2, \|\Theta\|_2 = [\theta_1, \theta_2, \ldots, \theta_m], \quad (1)$$

where the geodesic distance $d(S_1, S_2)$ was calculated based on the principal angles $[\theta_1, \theta_2, \ldots, \theta_m]$ between spaces (S_1, S_2), which were obtained from the orthonormal basis matrix (obtained by ONPE) and singular value decomposition.

$$score_{(x,i)} = \frac{exp(d_{(x,i)} - d_{max})}{\sum_{j=0}^{6}(d_{(x,j)} - d_{max})}, \quad (2)$$

$$d_{max} = max(d_{(x,0)}, d_{(x,1)}, \ldots, d_{(x,6)}), \quad (3)$$

$$d_{(x,i)} = -d(S_x, S_i), \quad (4)$$

where $score_{(x,i)}$ represents the similarity score between the testing data x and the deepfake category i, while $d_{(x,i)}(i = 0, 1, \ldots, 6)$ represents the negative of the geodesic distance between the testing data manifold space S_x and the deepfake method i manifold space S_i.

3.5. Model fusion

To effectively improve the final recognition results, we conducted model fusion at three levels.

3.5.1. Fusion on label layer

First is the label layer fusion. In the output scores of rawnet2, SE-Res2Net50 and HuBERT models, the index corresponding to the maximum score was set to be the output label. A threshold was set for open-set recognition based on model training and validation. The output labels were secondary adjusted and those with scores less than the threshold were considered as unknown label 7. Finally, three sets of recognition label values were thus obtained for the testing data. The mode of the three sets of labels was used as the fused label. When all three sets of labels were different, the result from HuBERT model was chosen as the fused result because it had the best performance.

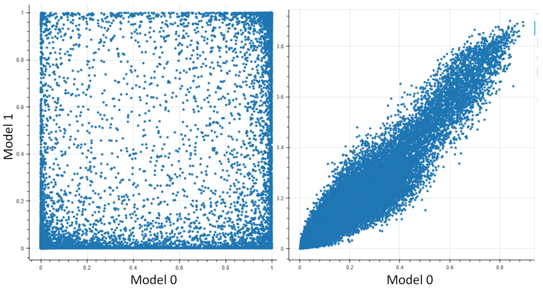

Figure 1: Score distribution for model0 and model1 (left: before inference augmentation, right: after inference augmentation).

3.5.2. Fusion on score layer

Next is the score-level fusion. A common score fusion method is conducted by calculating the mean of multiple sets of scores. As discussed in literature [9], when the scores showed a clear polarization in the histogram, it would be hard to perform score fusion, and the fusion results maybe degraded. In our work, the scores we obtained of the testing data showed a polarization in the histograms, as shown in Figure 1 (left). Although this phenomenon is not as prominent as in the literature [9], we had taken a measure of inference augmentation to alleviate it. As we know, if a model is trained well on the training set, the Softmax function will be likely to get extreme values (0 or 1). To make the outputs of softmax less close to 0 or 1, we first set a bound of (-20,20) and then added a constant multiplier of 0.1 to the inputs of softmax. The score distribution after inference augmentation is shown in Figure1 (right). Then the index corresponding to the maximum score was set to be the output label.

3.5.3. Fusion on feature layer

Finally, feature-level fusion is performed, as shown in Figure 2. For different models, the 256-dimensional output of the penultimate layer was connected as the embedding features, and then used to construct the manifold space for each class, and the spatial distance was calculated as the similarity score.

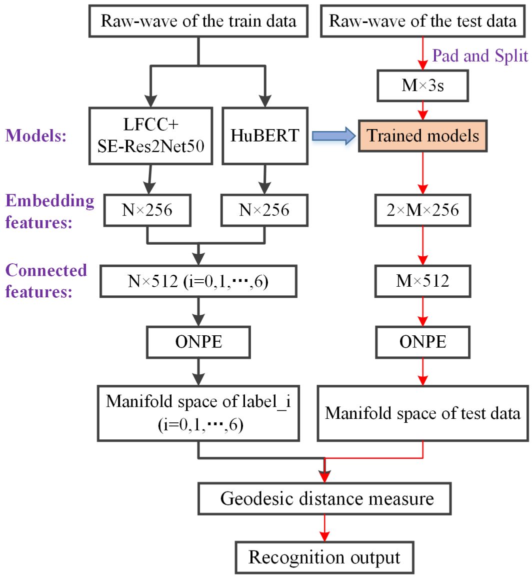

Figure 2: Overall framework of the fusion on feature layer.

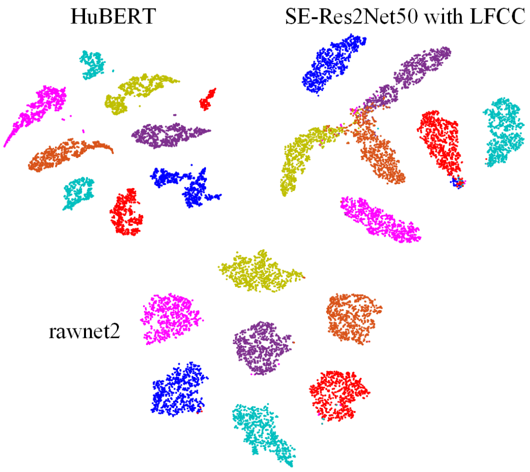

Figure 3: A t-SNE visual of different models. The dots of different colors represent different classes.

Specifically, N training data for each deepfake method i were input to the K trained models to obtain $N \times 256 \times K$ feature matrices. The feature matrices were then processed by ONPE to obtain the manifold space of the deepfake method. Next, the testing data were segmented into segments of length $3s$ with a shift of $1s$, and M audio segments were obtained. The segments were input to the three trained models to obtain $M \times 256 \times K$ feature matrices. The feature matrices were processed by ONPE to obtain the manifold space of the testing data. The geodesic distance and softmax score between manifold space of training data and manifold space of testing data were calculated as the final fusion score. If the maximum score was higher than the threshold, the index corresponding to the maximum score was set to be the output label, otherwise the label was set to 7 as a new label, and the threshold was fine-tuned by testing data.

4. Results and discussion

Table 1 shows the results of our proposed methods on ADD 2023 Track 3. We can find the following observations.

Firstly, the model of HuBERT works better than rawnet2 and Se-Res2Net50 with LFCC. The used chinese-hubert-large model is trained on large datasets, which truly helping reduce overfitting and improving the robustness of recognition results. Although at the beginning we thought rawnet2 should be able to achieve better results, in fact it was less effective than the approach of Se-Res2Net50 with LFCC, probably because the model was relatively simple and overfitting was more serious. Maybe we should match the model with appropriate subsequent classification networks so as to train a model with excellent discriminative ability. To visualize the effectiveness of the proposed method, we also used t-SNE [24] to visualize the embedding features of the three models on developing data, as shown in Figure 3. It can be seen that the distinguishability of rawnet2 was better than that of Se-Res2Net50 on developing data, which also indicated that the trained rawnet2 model was over-fitted from another perspective.

Secondly, in terms of fusion strategies, it can be seen that manifold-based feature-level fusion got the best performance, while the score-level fusion by inference augmentation performed better than common score fusion method (shown as F21 vs F22 and F41 vs F42). As our trained rawnet2 model got poor performance and it pulled down the overall performance in label-level fusion with the other two models (shown as F1), it was not considered in the subsequent score-level fusion and feature-level fusion.

According to the results of score-level fusion and feature-level fusion, it indicated that there was complementary information among the different models, and by constructing the manifold space and measuring the geodesic distance, further discriminative information was extracted, thus enhancing the overall recognition performance.

Thirdly, in data augmentation, shown as B11 to B22, due to the variability of background noise between the training and testing data, by adding noise to the training data was effective in improving the model performance.

Table 1
The results of F1-score for our proposed different methods on ADD 2023 Track3 (DA: data augmentation, FT: fine-tuned, LF: label fusion, SF: score fusion, IA: inference augmentation, FF: feature fusion)

NO.	Method	Add	F1-score
B11	Raw wave+rawnet2	-	0.5674
B12	Raw wave+rawnet3	DA	0.6470
B21	LFCC+Se-Res2Net50	-	0.6332
B22	LFCC+Se-Res2Net50	DA	0.7202
B31	HuBERT	-	0.7581
B32	HuBERT	DA	0.7238
F1	B12+B22+B32-LF	-	0.7121
F21	B22+B32-SF	FT	0.7280
F22	B22+B32-SF	IA+FT	0.7302
F3	B22+B32-FF	FT	0.7352
F41	B22+B31-SF	FT	0.7477
F42	B22+B31-SF	IA+FT	0.7608
F5	**B22+B31-FF**	**FT**	**0.7934**

However, unexpectedly, the performance of the HuBERT model trained on the augmented data was not as good as that of the HuBERT model trained on the original training data (shown as B31 vs B32). One possible reason was that the training data of the pre-trained models already contain rich noisy data, which itself can shield the effect of noise on speech. In addition, due to the time constraint of the competition, all models were obtained by training a set of parameters and no parameter tuning was performed, which may also be a reason.

Finally, it should be noted that, F3 in Table 1 with a F1-score of 0.7352, was the best result we submitted to ADD Track 3 during the competition and is ranked 5th. After the competition, when we conduct supplementary experiments on data augmentation, a better result was found as B31. Then we conduct relevant fusion experiments and obtained results shown as F4 and F5, with the best result up to 0.7934, which so far can rank 3rd in the competition. Despite this, the conclusion that feature-level fusion was better than fractional-level fusion was consistent.

5. Conclusion

The existing fake audio recognition systems often rely on three types of architectures: handcrafted features with classifiers, end-to-end classification models, and pre-trained feature extractors with classifiers. In ADD Track 3, we explored three models and three multi-model fusion strategies. Experiments demonstrated the effectiveness of the proposed manifold-based feature-level fusion strategy. And the proposed score-level fusion by inference augmentation provided an attempt to solve the fusion of models with an overfitting tendency. In addition, we experimented the effect of data augmentation on model performance enhancement. Finally, the proposed model fusion method obtained the F1-score of 0.7934 in ADD Track3 testing.

References

[1] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).

[2] X. Tan, T. Qin, F. Soong, T. Y. Liu, A survey on neural speech synthesis (2021).

[3] J. Muhammad, Akbar, A overview of spoof speech detection for automatic speaker verification (2019).

[4] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, H. Meng, Replay and synthetic speech detection with res2net architecture, in: International Conference on Acoustics, Speech, and Signal Processing, 2021.

[5] C. I. Lai, N. Chen, J. Villalba, N. Dehak, Assert: Anti-spoofing with squeeze-excitation and residual networks (2019).

[6] Z. Wu, R. K. Das, J. Yang, H. Li, Light convolutional neural network with feature genuinization for detection of synthetic speech attacks (2020).

[7] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, 2021, pp. 6369–6373. doi:10.1109/ICASSP39728.2021.9414234.

[8] X. Liu, M. Liu, L. Zhang, L. Zhang, C. Zeng, L. Kai, N. Li, K. A. Lee, L. Wang, J. Dang, Deep spectro-temporal artifacts for detecting synthesized speech (2022). doi:10.48550/arXiv.2210.05254.

[9] Y. Zhang, J. Lu, X. Wang, Z. Li, R. Xiao, W. Wang, M. Li, P. Zhang, Deepfake detection system for the add challenge track 3.2 based on score fusion, 2022, pp. 43–52. doi:10.1145/3552466.3556528.

[10] C. Geng, S.-J. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021) 3614–3631. doi:10.1109/TPAMI.2020.2981604.

[11] A. Bendale, T. E. Boult, Towards open set deep networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1563–1572. doi:10.1109/CVPR.2016.173.

[12] C. Ding, K. Liu, F. Cheng, E. Belyaev, Spatio-temporal attention on manifold space for 3d human action recognition, Applied Intelligence 51 (2021). doi:10.1007/s10489-020-01803-3.

[13] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang,

X. Yan, L. Xu, Z. Wen, H. Li, Add 2022: the first audio deep synthesis detection challenge, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 9216–9220. doi:10.1109/ICASSP43922.2022.9746939.

[14] D. Snyder, G. Chen, D. Povey, Musan: A music, speech, and noise corpus (2015).

[15] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, CoRR abs/2006.11477 (2020). URL: https://arxiv.org/abs/2006.11477. arXiv:2006.11477.

[16] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, CoRR abs/2106.07447 (2021). URL: https://arxiv.org/abs/2106.07447. arXiv:2106.07447.

[17] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, F. Wei, Wavlm: Large-scale self-supervised pre-training for full stack speech processing, CoRR abs/2110.13900 (2021). URL: https://arxiv.org/abs/2110.13900. arXiv:2110.13900.

[18] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, rawnet2-antispoofing (2021). URL: https://github.com/eurecom-asp/rawnet2-antispoofing.

[19] X. Li, N. Li, C. Weng, et al., asv-anti-spoofing-with-res2net (2020). URL: https://github.com/lixucuhk/ASV-anti-spoofing-with-Res2Net.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[21] C. Peng, S. Liu, chinese-speech-pretrain (2022). URL: https://github.com/TencentGameMate/chinese_speech_pretrain.

[22] X. Liu, J. Yin, Z. Feng, J. Dong, L. Wang, Orthogonal neighborhood preserving embedding for face recognition, in: 2007 IEEE International Conference on Image Processing, volume 1, 2007, pp. I – 133–I – 136. doi:10.1109/ICIP.2007.4378909.

[23] T. Wang, P. Shi, Kernel grassmannian distances and discriminant analysis for face recognition from image sets, Pattern Recognition Letters 30 (2009) 1161–1165. URL: https://www.sciencedirect.com/science/article/pii/S0167865509001391. doi:https://doi.org/10.1016/j.patrec.2009.06.002.

[24] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.