Singing Voice DeepFake Detection

Mahyar Gohari[1], Davide Salvi[2], Paolo Bestagini[2], Nicola Adami[1]
[1]Department of Information Engineering, University of Brescia, Italy
[2]Department of Electronics, Information and Bioengineering, Polytechnic University of Milan, Italy

November 6, 2024

This document describes a system designed to effectively detect deepfake-generated utterances of singing voices in real-world scenarios. The paper is organized as follows: Section 1 details the data pre-processing steps. Section 2 outlines the employed methodology. Section 3 describes the training process and parameter choices. Finally, Section 4 presents the conclusions drawn from this work.

1 Data Preprocessing

The system begins with the segmented WildSVDD dataset, from which we process the audio tracks to ensure all segments have a uniform length of 4 seconds. This involves segmenting tracks into multiple parts, each of 4 seconds or less, and padding those segments that are shorter than 4 seconds. It is noteworthy that no data augmentation techniques were applied in this study.

2 Methodology

For our system, we utilize the log-spectrogram representation of audio tracks, which are fed into the ResNet18 model [2], serving as the backbone of our architecture.

The log-spectrograms are computed from audio tracks sampled at 44.1 kHz using a window length of 25 ms and a hop size of 10 ms, resulting in 1024 frequency bins. To enhance performance, the log-spectrograms are normalized using the same parameters as the ImageNet dataset.

These log-spectrograms are then input into the standard ResNet18 architecture, which has been pretrained on ImageNet [1]. We modified the weights of the initial layer to accommodate the input shape by averaging across the three color channels. Additionally, we replaced the final fully connected layer of the pre-trained model to adapt it to our binary classification task. The new head includes a dropout layer with a probability of 0.5 to mitigate overfitting, followed by a 256-unit linear layer with ReLU activation to capture task-specific features. A final single-output linear unit generates the predicted score for binary classification.

During inference, to estimate the scores for each segment in the original dataset, we averaged the predicted scores of all 4-second segments.

3 Training

We trained the model twice: first on mixed songs and then on separated vocals. The model trained on mixed songs was used to predict results for the test dataset containing mixed songs, while the model trained on separated vocals was used to predict results for the test dataset containing separated vocals. We trained the models for 100 epochs with a batch size of 64 samples.

To facilitate learning from difficult samples, we employed sigmoid focal loss with a focusing parameter $\gamma = 2$. The learning rate followed a cosine annealing schedule, starting at 10^{-4} and decaying to 10^{-7} over the course of training.

Additionally, we applied a weight decay of 10^{-4} for regularization, selecting the best-performing model based on the lowest validation loss.

To address potential class imbalance in the training batches, we utilized random over-sampling to ensure each batch contained an equal number of samples from both classes.

4 Conclusions

In conclusion, the proposed system demonstrates effective detection of deepfake singing voices by leveraging a robust architecture and a carefully crafted training regimen. Future work may involve exploring advanced augmentation techniques and the application of more complex model architectures to further improve detection performance.

Acknowledgments

This work was carried out as part of the 37th cycle PhD program in Information Engineering at the University of Brescia. The research was partially funded by Regione Lombardia through the initiative "Programma degli interventi per la ripresa economica: sviluppo di nuovi accordi di collaborazione con le università per la ricerca, l'innovazione e il trasferimento tecnologico" - DGR n. XI/4445/2021.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.