CNN AND MKDE-BASED CLASSIFICATION OF SYNTHETIC SPEECH ATTRIBUTION

TCLAB: Youngjun Sim, Sechan Oh, Hyungsup Yoon, Jungyu Choi, and Sungbin Im

School of Electronics, Soongsil University, Seoul, Korea

ABSTRACT

Speech synthesis algorithms developed over the past few years can be easily used by the general public and have excellent performance. If these technologies are used maliciously, they can be applied to various crimes such as people's impersonation and fake news. For this reason, many studies have been conducted recently to solve these problems, and various synthetic speech detectors have been developed in many previous studies. In this study, we propose a synthetic speech audio detection model capable of open-set classification even in the presence of noise. The proposed approach consists of convolutional neural network and multivariate kernel density estimation. In Path 1, the mel spectrogram of the audio signal is used as a feature, and in Path 2, the kernel density function of the training class is estimated using the logit value of CNN as a feature. The data used for model training and evaluation are provided at the 2022 IEEE Signal Processing Cup. The accuracy of the final model is 96.5 % and 95.5 % for the evaluation set of Part 1 and Part 2 of the competition, respectively.

Index Terms— Speech Synthesis, Melspectrogram, Convolution Neural Network, Multivariate Kernel Density Estimation

1. INTRODUCTION

Recently, various speech processing algorithms such as speech conversion or speech synthesis have been developed. Representative examples include Wavenet[1] , WORLD vocoder[2], Merlin[3], MaryTTS platform[3], and auto-encoder[4]. These speech synthesis algorithms can be easily used by ordinary people and their performance is excellent. These advances imply that more tools are available to digital artists working with voice. When these schemes are used maliciously, including distributing fake news impersonating celebrities or illegally authorizing user access to the automated speaker verification system (ASV), it can cause serious problems [5].

Therefore, many studies are being conducted to solve these problems, and various synthetic voice detectors have been developed in many previous studies to detect spoofing techniques that attack the ASV system. Reference [5] proposes a latent profile analysis (LPA) classifier capable of open-set classification with the ASVspoof2015 dataset. In reference [6], using a distorted version of the ASVspoof 2015 dataset, a framework is put forward that allows closed-set classification even when reverberation and noise are added. However, there are few prior studies on open set classification as synthetic voice samples corrupted by noise.

In this report, we propose a synthetic speech audio detection model capable of open-set classification even in the presence of noise. This model classifies voice inputs according to the algorithms used to synthesize them. Three scenarios are assumed to implement the proposed model: closed-set, open-set, and data augmentation. In the closed-set scenario, the classifier classifies which algorithm is used for speech generation within the known algorithms. In the open-set scenario, the unknown algorithm is included in addition to the known algorithms. The classifier determines whether or not the synthesized speech input belongs to the unknown algorithm. If the speech input falls into the known algorithm group, it can detect its type. The model is extended in the data augmentation scenario so that open-set classification is possible even in synthetic voice samples affected by noise, reverberation, and compression. The proposed approach is developed based on convolutional neural network (CNN) and multivariate kernel density estimation (MKDE). The mel spectrogram of the audio signal is used as a feature for CNN. In MKDE, the kernel density function of the training class is estimated using the logit values of CNN as features.

The structure of this report is as follows. Section 2 describes the technique used in this experiment as the proposed method. The proposed one includes the convolutional neural network (CNN) model used, including feature extraction and the multivariate kernel density estimation (MKDE) for final voting. Section 3 presents the data used in the experiment. The results of the experiments are summarized in Section 4, and the conclusion is given in Section 5.

2. PROPOSED METHOD

2.1. Preprocessing

In this section, we introduce the proposed approach for detecting synthesized speech for open set recognition. The proposed one takes the Mel spectra of an audio signal file as feature input. Since the lengths of the audio files are not regular, the proposed approach employs the data slicing technique, in

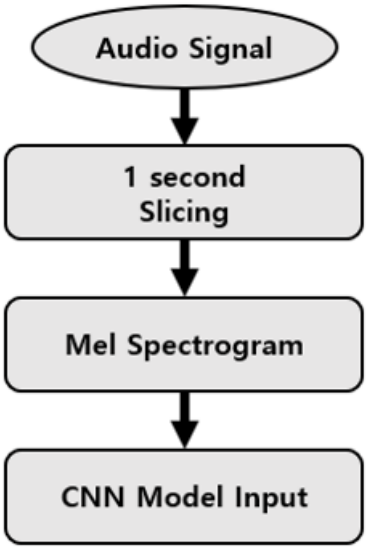

Fig. 1. Flow chart of feature generation process with an audio file.

which an audio file is sliced at 1-second intervals to create fixed-length speech segments. For example, if an audio file with a length of 10.5 seconds is cut at 1-second intervals, a total of 10 slices are created from 0 seconds to 10 seconds, and the last 0.5-second samples are disregarded. After slicing, a Mel spectrogram is obtained for each slice and used as a feature of the proposed model. Fig. 1 summarizes this process of generating the features from an audio file.

Fourier transform expresses an arbitrary time-domain signal as the sum of periodic functions having various frequencies. However, the time dependency of the frequency components disappears in the Fourier transform. In order to compensate for this disadvantage, the short-time Fourier transform (STFT) is employed, which relies on the windowing process. STFT segments a signal into window-length units, and Fourier transform is applied to each frame [7].

Mel spectrogram is a conversion of the STFT-based spectrogram components to the mel scale tailored to human auditory characteristics. Below 1 kHz, the center frequency is evenly divided, and above 1 kHz, it follows the logarithmic scale. The mel scale conversion formula [8] for the frequency f is given by

$$Mel(f) = 2595log(1 + \frac{f}{700}) \quad (1)$$

Mel-filter bank refers to a triangular filter implemented by approximating the mel-scale function through filter banks in the frequency domain. It responds according to a linear scale to a signal in the low-frequency region but has a logarithmic scale characteristic in the high-frequency region. The mel spectrogram based on the auditory structure showed good performance in the deep learning model dealing with speech data [9, 10]. For this reason, the mel spectrogram is adopted as a feature extraction technique for the proposed model.

In the experiment, we compute the mel spectrograms of audio slices with a sampling rate of 16 kHz and a length of 1 second for features. This process utilizes 2048-sample windowing and 25% overlapping to make a total of 32 frames for one slice. Hamming window is applied to minimize the spectral leakage phenomenon [11]. The number of bins of the mel scale is set to 64, and the computed mel spectrogram is grouped into 64(bin)×32(frame) because the feature size for the CNN model in this study is set to 64 $\times$32. Fig. 2 shows the mel spectrogram examples of the provided known and unknown audio signals.

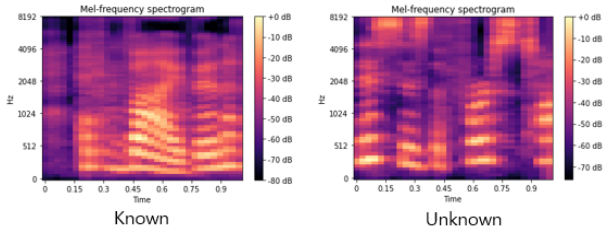

Fig. 2. Mel Spectrogram examples of Known and Unknown audio signals.

2.2. Proposed Approach

Fig. 3 shows the flow chart of the class prediction process employed in this study. In the figure, the mel spectrogram feature is input to the CNN model, and the model outputs the logit value corresponding to the feature. The logit value is fed into Paths 1 and 2, respectively. In Path 1, the softmax function carries out the label prediction denoted by y_1 for the logit value. At the same time, the MKDE of Path 2 produces the kernel density y_2 corresponding to the input logit value, and y_2 is compared with the predetermined threshold value THX. If $y_2 < THX$, the feature belongs to the unknown algorithm group. Otherwise, it belongs to the known algorithm group, and its predicted label is determined according to y_1.

2.2.1. CNN Model

Fig. 4 displays the CNN structure employed in this competition.For the structure of the neural network, BatchNormalization, Conv2D, and ReLU are repeatedly connected, and MaxPool2D is connected. BatchNormalization is used as the first layer of the model to make learning faster through regularization. After BatchNormalization, 16 filters of 5x5 size are used, and a Conv2D layer employing padding is connected so that the input and output sizes are the same. In addition, to solve the gradient vanishing problem, the output of the Conv2D layer is passed through the ReLU activation function. This structure is repeated two more times by increasing only the number of filters of Conv2D to 32 and 64. After that,

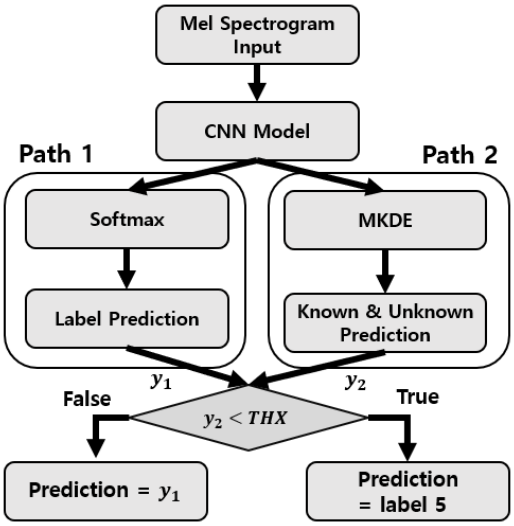

Fig. 3. Flow chart of the proposed classification procedure.

feature information is compressed by attaching one MaxPooling2D layer with a pooling filter size of 5x5.

Finally, in the fully connected layer responsible for classification, the number of output layer nodes is set to six to classify five known and one unknown algorithms, and a softmax activation function is applied. After that, the class of the instance is predicted by soft voting on the softmax values of the slices.

2.2.2. MKDE Algorithm

Multivariate kernel density estimation (MKDE) is a non-parametric density estimation technique that estimates a multivariate kernel density function using only observed data with a given kernel function [12, 13]. A kernel function is defined as a symmetric non-negative function about the origin and whose integral value is 1. The MKDE model in this study determines the unknown algorithm class that the CNN model cannot identify.

As the input data of the MKDE model, robust scaling is applied to the logit value output by the CNN model. The Wang-Ryzin kernel function that gives the highest performance through the experiment is employed. Bandwidth is calculated by Silverman's rule and is tuned through experiments. Using the density function estimated with the training data, the kernel density value of the logit value of the input instance is computed and the MKDE model identifies instances lower than the threshold value as an unknown algorithm class. Among the instances classified in Path 1, the instances identified as an unknown algorithm group by the MKDE model are finally classified as the unknown algorithm group.

3. DATASET DESCRIPTION

The dataset used is from the IEEE Signal Processing Society's 2022 Signal Processing Cup data. This dataset includes training data generated by five known algorithms and unseen data which belong to an unknown algorithm. For the training data, a total of 5,000 voice files are provided, with the files labeled from 0 to 4, and they are one thousand for each label. Additional one thousand audio files are given as unseen data, and we classified them as label five and used them in this experiment. The training and unseen data provided in the competition are given in wav file format. The length of the audio file varies from 1.35 to 14.77 seconds for the training data and 1.57 seconds to 11.61 seconds for unseen data. The speakers of the audio file are male and female, the language used is English, and English pronunciations of various nationalities are included.

To obtain distorted data to be used in the competition part 2, a Matlab script is used for noise, reverberation, and compression files. We apply this script to additionally create 15,000 voice files with noise derived from the train data and 3,000 voice files with noise from the unseen data. In this process, the files contaminated by noise and reverberation are in a wav file format, while the files by compression use FFmpeg to create an mp3 file. We convert mp3 files into Wav files to make it easier to handle audio files. In this way, we use a total of 24,000 voice files consisting of 20,000 training data with or without noise and 4,000 unseen data with or without noise.

4. EXPERIMENTS

4.1. Open Competition Part 1

The number of training data for training the CNN model in Path 1 is 15,360 out of a total of 24,000 files and 3840 files are used for validation. The remaining 4,800 files are used for testing, including 1,238 files without distortion. In the MKDE model of Path 2, the kernel density is estimated using 4,000 files of the known algorithm group that is without noise. The kernel density values of the 1,000 remaining files from the known algorithm group and the 1,000 files of the unknown algorithm group are shown in Fig. 5. File numbers from 1 to 1,000 are from the known algorithm group, and file numbers from 1,001 to 2,000 represent the unknown algorithm group. The optimal threshold value is determined using the evaluation set of the competition. Finally, among the predictions of the CNN model of Path 1, instances with a kernel density less than the threshold value are reclassified as label five.

4.2. Open Competition Part 2

The training, validation, and test data for the CNN model of Part 2 are the same as those of Part 1. In the MKDE model, the kernel density is estimated using 20,000 known algorithm files with noise, and the kernel density values of 4,000 known

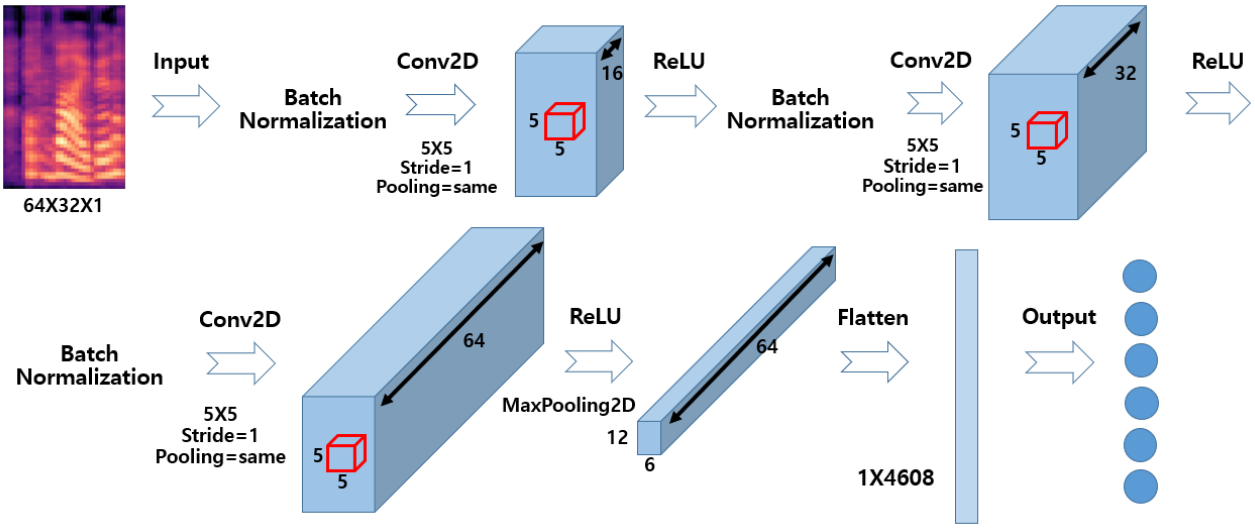

Fig. 4. CNN architecture employed in this study.

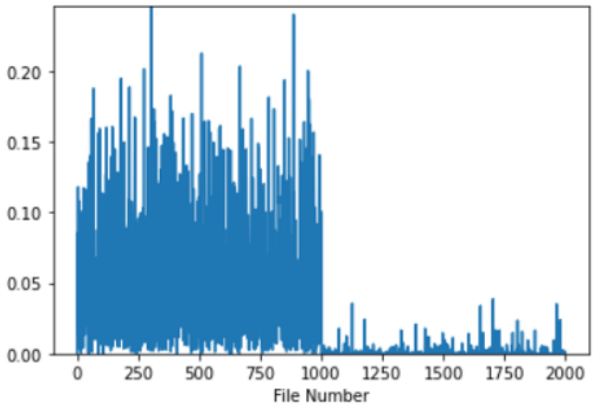

Fig. 5. Kernel density of the test data for Part 1.

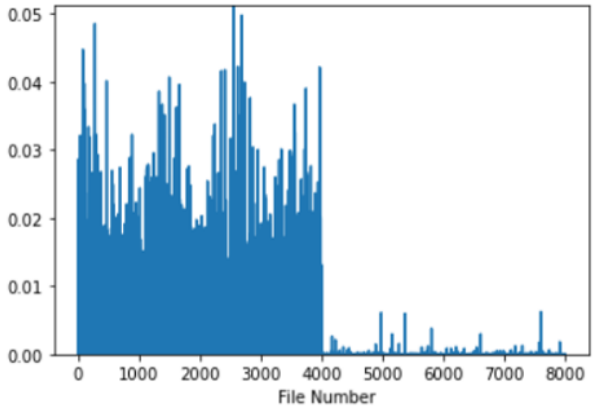

Fig. 6. Kernel density of the test data for Part 2.

algorithm files and 4,000 unknown algorithm files are shown in Fig. 6. File numbers from 1 to 4,000 represent the known algorithm files, and file numbers from 4,001 to 8,000 do the unknown algorithm files. The optimal threshold value is determined using the evaluation set of the competition.

5. RESULT AND CONCLUSION

In this report, we proposed a novel method to detect which algorithm had been used to generate the synthetic speech tracks from the IEEE Signal Processing Cup 2022 open competition in an open set. The proposed method is based on a combination of supervised learning and unsupervised learning approaches. A set of mel spectrogram features is extracted from the audio. The CNN-based supervised classifier is trained to solve the classification problem. In contrast, the unsupervised classifier based on MKDE supports recognizing whether the predicted data belongs to the known or unknown algorithm category. For the evaluation dataset given for the open competition, our proposed detector demonstrates the accuracy of 96.5 % in Part 1 and 95.5 % in Part 2. This performance of the proposed one is outstanding compared to other open-set recognition methods that we tried. Despite the promising results, we believe that there can be more improvement considering other conditions.

6. REFERENCES

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu,

"Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[3] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system.," in *SSW*, 2016, pp. 202–207.

[4] Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner, "Open source voice creation toolkit for the mary tts platform," in *12th Annual Conference of the International Speech Communication Association-Interspeech 2011*. ISCA, 2011, pp. 3253–3256.

[5] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, pp. 1–14, 2021.

[6] Yanmin Qian, Nanxin Chen, Heinrich Dinkel, and Zhizheng Wu, "Deep feature engineering for noise robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1942–1955, 2017.

[7] Roneel V Sharan and Tom J Moir, "Acoustic event recognition using cochleagram image and convolutional neural networks," *Applied Acoustics*, vol. 148, pp. 62–66, 2019.

[8] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125868–125881, 2019.

[9] David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive, "Ina's mirex 2018 music and speech detection system," *Music Information Retrieval Evaluation eXchange (MIREX 2018)*, 2018.

[10] Pooyan Mobtahej, Xulong Zhang, Maryam Hamidi, and Jing Zhang, "Deep learning-based anomaly detection for compressors using audio data," in *2021 Annual Reliability and Maintainability Symposium (RAMS)*. IEEE, 2021, pp. 1–7.

[11] Jia-Shing Sheu and Ching-Wen Chen, "Voice recognition and marking using mel-frequency cepstral coefficients," *Sensors and Materials*, vol. 32, no. 10, pp. 3209–3220, 2020.

[12] Hanzi Wang and David Suter, "Mdpe: A very robust estimator for model fitting and range image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 139–166, 2004.

[13] Hanzi Wang, "Maximum kernel density estimator for robust fitting," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3385–3388.