Fake Speech Answer-Sheet Detection on Intelligent Learning App Based on Block-Based Deep Neural Network

Qingzhu Wu¹, Haitao Guo²,* and Shaowei Xiong³

¹Guangdong Mechanical and Electrical Polytechnic, Guangzhou, China
²Guangdong Construction Polytechnic, Guangzhou, China
³Guangdong Mechanical and Electrical Polytechnic, Guangzhou, China

*Corresponding author: guohaitao@gdcvi.net

Abstract. In order to detect the fake speech answer-sheet obtained by text-to-speech or voice conversion algorithms on intelligent English learning app, this paper proposed the idea of block-based deep neural network (BDNN), which is constructed to extract deep features. It is composed of classification-based blocks. In BDNNs, the deep feature extracted by a preceding block serves as the input to its following block. Different from the commonly used inputs for deep features extraction, the magnitude-phase spectrum, including magnitude and phase information, is exploited to feed BDNNs in order to extract effective deep features. The experimental results show that the proposed method can detect most of fake speech answer-sheet on intelligent English learning app.

Keywords: Fake speech answer-sheet, detection, intelligent language learning app, block-based neural network.

1. Introduction

In recent years, artificial intelligence (AI) technology is applied more and more in language teaching. In China, IFLYTEK developed idea future smart teaching platform (FiFSTP) based on AI, which contains oral English teaching, learning and examination modules. Today, more than 1,000 universities or collages have used it for courses teaching.

One of the important functions of FiFSTP is oral English learning. In the process of learning, aiming at the characteristics of Chinese students who do not like to speak English, the system firstly gives the text of utterance or paragraph. If the student wants to listen to the correct pronunciation, he can listen after pressing the corresponding button. In addition, after the student reads the utterance or paragraph and submits the recording, the system can give the score according to the pronunciation, fluency and completion for every utterance by the users, which can help the students improve oral English skills.

However, there exist some problems when using FiFSTP as a teaching app. When teachers give oral English assignments, especially examination on the app, some students may use other tools or voice recording to cheat the system in order to get a higher score. For example, it is easy to convert the text into speech by using a toolkit of text-to-speech [1] or voice conversion [2] and hand in the converted speech as the speech answer-sheet.
In order to be fair and seek the students who fool the system by using the toolkit of text-to-speech or voice conversion algorithms, it is necessary to develop a system to detect the fake speech answer-sheet on the intelligent oral English system.

2. The proposed method

The same as many speech information processing systems, the proposed system to detect fake answer-sheet in oral English contains two parts: front-end feature extraction and back-end classification. In which, block-based deep neural network (BDNN) is used to extract deep feature and DNN is used as classifier. In addition, magnitude-phase spectrum (MPS) [3] is used as the input of BDNN because it has magnitude information and phase information. Next, we will introduce them one by one.

2.1. Magnitude-phase spectrum

Because MPS is extracted on the basis of constant-Q transform (CQT), we introduce CQT first and then MPS.

2.1.1. Constant-Q transform. Given a speech signal \( x(b) \), its CQT \( X(k, b) \) can be written as [4]:

\[
X(k, b) = \sum_{j=b}^{j+\left\lfloor \frac{B_k}{2} \right\rfloor} x(j) a^*_k (j - b - \frac{B_k}{2}), k = 1, 2, ..., K
\]

Where \( K \) is the number of frequency bin indexes, \([\cdot]\) and \(*\) denote the great integer function and the complex conjugate operator, respectively. \( B_k \) is the variable windows length and the basic function \( a_k(b) \) represents the complex-valued time-frequency atom, given by

\[
a_k(b) = \frac{1}{C} \nu \left( \frac{b}{B_k} \right) \exp \left[ i(2\pi b \frac{f_k}{f_s} + \phi_k) \right]
\]

Where \( f_k = f_s \frac{k-1}{B_{num}} \) is the centre frequency of the \( k \)-th bin, \( f_i, f_s \) and \( \phi_k \) are the centre frequency of the lowest-frequency bin, the sampling rate and the phase offset, respectively. \( \nu(\cdot) \) denotes a window function. \( B_{num} \) and \( C \) are the number of bins in every octave and a scaling factor, respectively, where \( C \) is defined as:

\[
C = \sum_{d=-\left\lfloor \frac{B_k}{2} \right\rfloor}^{\left\lfloor \frac{B_k}{2} \right\rfloor} \nu \left( \frac{d + \frac{B_k}{2}}{B_k} \right)
\]

where \( d \) represents the sample number in the window.

2.1.2. Magnitude-phase spectrum. \( X(k, b) \) in Eq. (1) can be rewritten as:

\[
X(k, b) = X(\omega) = |X(\omega)| e^{i\phi(\omega)}
\]
Where \( \omega = 1, 2, ..., K \) is the frequency bin index, \(|X(\omega)|\) and \(\varphi(\omega)\) stand for magnitude and phase spectra of \(x(b)\), respectively. Taking the natural logarithm function of Eq. (4), we obtain:

\[
\ln(X(\omega)) = \ln(|X(\omega)|e^{j\varphi(\omega)}) = \ln(|X(\omega)|) + j\varphi(\omega) \ln(e)
\]

\[
= \ln(|X(\omega)|) + j\varphi(\omega)
\]

(5)

Where \(\ln(\bullet)\) denotes the natural logarithm function.

The modulus of Eq. (5) is:

\[
|\ln(X(\omega))| = |\ln(|X(\omega)|) + j\varphi(\omega)|
\]

\[
= \sqrt{\left(\ln(|X(\omega)|)\right)^2 + (\varphi(\omega))^2}
\]

(6)

Where \(\ln(X(\omega))\) denotes the MPS of \(x(b)\). From Eq. (6), it can be seen that both magnitude information and phase information are including in MPS.

2.2. Block-based deep neural network

The proposed BDNN is constituted by several block. Here, we introduce block first and then BDNN.

DNN is commonly used in classification. Hence, we construct block by DNN. The framework of a block is shown in Figure 1. From Figure 1, it can be seen that the block is composed of \(D + 2(D > 1)\) layers, namely, one slice layer, \(D\) fully connected layers (also named as hidden layers) and one output layer. Each layer contributes to different functions:

- Slice layer performs a function of receiving context and current information, for example by using 11 frames combined with 5 left frames, 5 right frames and 1 current frame.
- Fully connected layers serve the function of feature learning.
- Output layer has respectively the classification function.

![Figure 1. The framework of a block](image)

The schematic diagram of deep feature based on MPS and BDNN is shown in Figure 2. From Figure 2, we observe that the BDNN is composed of \(n\) blocks, where each block generates its deep
feature $D_{\text{fae}_m} (1 \leq m < n)$ and passes it to the next, i.e., $(m+1)$-th block. For every block training, the same training data corresponding to label information is used. In addition, the blocks are trained in ascending order. That's to say, $i$-th block begins to be trained after $(i-1)$-th block training is finished, $i = 1, 2, ..., n$.

Now we compare the context information between traditional DNN and the proposed BDNN. We assume that the current frame is $t$, the context information in the 1st block is from $t-5$ to $t+5$, followed by that in the 2nd block from $t-10$ to $t+10$, and from $t-15$ to $t+15$ in the 3rd block, respectively. Similarly, the context information of the $i$-th block is from $t-5i$ to $t+5i$. Therefore, we can see that more context information is used by BDNNs if more blocks are used. However, the context information in conventional DNN is usually from $t-5$ to $t+5$.

![Diagram](image.png)

**Figure 2.** Schematic diagrams of deep feature extraction based on MPS and BDNN.

Following the previous works on DNN framework [5], we utilize the following rules to train every block:

1. Sigmoid activation function is used to train the hidden layers.
2. Cross-entropy with softmax is utilized as training criterion for classification layer of the classification-based blocks.
3. Mean and variance normalization are applied to the input data for better performance.
4. Stochastic gradient descent is exploited as the optimization method during training process.
5. For every block training, there are total 25 epochs, including the 1st epoch with 0.8 learning rate, 14 epochs from 2nd to 15-th epochs with 3.2 learning rate, and the rest with the learning rate of 0.08. Regarding the first epoch, the minibatch size of 256 is used and the remaining epochs have a minibatch size of 1024 with a momentum value 0.9.

2.3. **Classifier**

In this work, DNN is used as the classifier for the deep feature. It has the same architecture with the block introduced in Figure 1.
3. Experiments and evaluation

3.1. Database and evaluation rule
due to the cost of data collection, there have not been publicly available database dedicated for fake answer-sheet detection in oral English examination. We collected 6318 genuine answer-sheet utterances of oral English examination from the FiFSTP by splitting and merging operation, the length of the utterances varies from 3s to 10s. They were divided into two parts randomly, part 1 has 2,200 utterances while part 2 has 4,118 utterances. In addition, 22,800 and 63,882 fake speech utterances obtained by different text-to-speech or voice conversion algorithms from ASVspoof 2019 logical access training set and evaluation set [7] were used to combine part1 and part 2 to from new training set (25,000=2,200+22,800) and test sets (68,000=4,118+63,882), respectively. In the area of fake answer-sheet detection, equal error rate (EER) can be considered as the performance metric, which denotes the error rate when the false rate $P_{FA}$ equals the miss rate $P_{MR}$. Here

$$P_{FA}(\lambda) = \frac{\#\{\text{Fake trials with score} > \lambda\}}{\#\{\text{Total fake trials}\}}$$ (7)

$$P_{MR}(\lambda) = \frac{\#\{\text{Genuine trials with score} \leq \lambda\}}{\#\{\text{Total genuine trials}\}}$$ (8)

Where $\lambda$ is a threshold.

3.2. Experimental setup
All the parameters in CQT are chosen the same as in [6]. For instance, the number of octaves $O$ and frequency bins ($B_{num}$) in every octave are set as 9 and 96, respectively, and the static dimension of MPS is 863. In addition, the structure of every block and the DNN-based classifier is set as 863x11:1024:1024:2.

3.3. Experimental results and analysis
The experimental result is given in Table 1. From Table 1, it can be seen that the EER of $Dfea_2$ can be reduced by 16.40% comparing with that of $Dfea_1$, which indicates that more discriminative information can be obtained from the second block on test set. however, $Dfea_3$ performs much worse than $Dfea_2$ on test set with the EER increasing from 6.78% to 10.22%. Moreover, $Dfea_1$ also outperforms $Dfea_3$ in terms of EER. The main reason is that overfitting has more significant impact on the performance.

| Deep feature | #Block | EER   |
|--------------|--------|-------|
| $Dfea_1$     | 1      | 8.11  |
| $Dfea_2$     | 2      | 6.78  |
| $Dfea_3$     | 3      | 10.22 |

| All $Dfea_m$ |
4. Conclusions
In this paper, we propose data-driven deep feature extraction method for speech answer-sheet detection on intelligent oral English training system. It is extracted from BDNN. The BDNN is composed of classification-based blocks. In addition, MPS with magnitude and phase information are applied to feed the BDNN. The experimental result demonstrates that BDNN can capture most of fake speech answer-sheets on the intelligent language learning app.

Acknowledgments
This paper is one of the phased research results of Guangdong Provincial Education Science "13th Five-Year" Plan 2020 Annual Research Project(2020GXJK577).

References
[1] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, “Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis,” IEEE/ACM Trans. Audio, Speech Lang. Process., 29 (2021), 274–285.
[2] B. Sisman, M. Zhang, and H. Li, “Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion,” IEEE/ACM Trans. Audio, Speech Lang. Process., 27 (2019), 1085–1097.
[3] J. Yang and L. Liu, “Playback speech detection based on magnitude-phase spectrum,” Electronics Letters, 54 (2018), 901–903.
[4] J. C. Brown, “Calculation of a constant Q spectral transform,” Journal of Acoustical Society of America, 89 (1991), 425–434.
[5] J. Yang and R. K. Das, and Haizhou Li, “Significance of subband features for synthetic speech detection,” IEEE Transactions on Information Forensics and Security, 15(2020), 2160–2170. 2020.
[6] M. Todisco, H. Delgado and N. Evans, “Constant Q cepstral coefficients: A spoofing countermeasure for automatic Speaker verification,” Computer Speech & Language, 45 (2017), 516–535.
[7] M. Todisco, X. Wang, V. Vestman, etc., “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” in 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 2019, pp. 1008–1012.