DETECTION OF VOICE TRANSFORMATION SPOOFING BASED ON DENSE CONVOLUTIONAL NETWORK

Yong Wang, Zhuoyi Su

School of Electronic and Information Engineering
Guangdong Polytechnic Normal University
GuangZhou 510006, China

ABSTRACT

Nowadays, speech spoofing is so common that it presents a great challenge to social security. Thus, it is of great significance to recognize a spoofed speech from a genuine one. Most of the current researches have focused on voice conversion (VC), synthesis and recapture which mimic a target speaker to break through ASV systems by increased false acceptance rates. However, there exists another type of spoofing, voice transformation (VT), that transforms a speech signal without a target in order 'not to be recognized' by increased false reject rates. VT has received much less attention. Thus, in this paper, we investigate the model of VT and propose a method using a very deep dense convolutional network with 135 layers to detect VT spoofed speeches from genuine speeches. The experimental results show that the average accuracies over intra-database and cross-database outperform the reported state-of-the-art methods.

Index Terms— spoofed speech, voice transformation, Dense Convolutional Network.

1. INTRODUCTION

Speech spoofing can be classified into two categories: 1) voice conversion (VC), synthesis and recapture that mimic a target speaker; 2) voice transformation (VT) that changes one's speech signal without a target [1]. They present different threats to security. The first ones are to change or capture one's speech, or to create an artificial speech, in order to be recognized as target person, while the second one is to change one's speech in order not to be recognized.

In recent years, reported efforts mainly focus on the detection of the first category. In [2] [3], long-term spectral statistics are used as the features and linear classfier is as the back-end classifier. In [4] [5] [6], mel-frequency cepstral coefficients (MFCC) are used as the features and GMM, SVM are as the back-end classifiers. In [7] [8], modified group delay is used as the features and HMM, GMM and SVM are as the back-end classifiers. In [9] [10], spectrogram is used as the input of the deep neural network (DNN) to recognize spoofed speeches.

It should be noted that since abundant information of target person is generally needed for VC and synthesis, and the situation for recapture is uncertain, there exist difficulties and costs in implementing the first category of spoofing to some extent. By contrast, no extra information is needed for VT spoofing implementation, leading to the fact that it has been integrated into many prevailing audio/speech editing tools while the first category has not, and it has been conducted in many criminal cases. However, compared with the researches on the first category detection, the researches on VT spoofing detection are relatively fewer and insufficient. Wu et al. and Wang et al. proposed algorithms of VT spoofing detection that employs MFCC statistics as features and SVM as classifier. Liang et al. proposed an approach based on convolutional neural network (CNN) for VT spoofing detection. The accuracies of the above methods are all less than 95%, indicating that an improvement is needed for practical applications.

Meanwhile, there has been a trend to use deep neural network as back-end classifier in spoofing forensics in recent year. The advantages of DNN framework over traditional classifiers such SVM or GMM, is that DNN can automatically extract deep features other than hand-designed features. In DNN, degradation occurs with the increment of network depth. In order to solve this problem, several solutions have been proposed among which is the Dense Convolutional Network (DenseNet). A DenseNet can significantly reduce the number of parameters, eliminate the need to relearn redundant feature-maps, strengthen the propagation feature, supports limited neuronal reuse and facilitates data training.

Therefore, in this paper, we examine the model of VT spoofing and propose an effective method of VT spoofing detection that employs spectrogram features and a very deep

This work was supported by the National Natural Science Foundation of China (61672173), the Characteristic Innovation Project of Guangdong Province Ordinary University (2015KTSCX083), the Natural Science Foundation of Guangdong Province(2014A030313623) and the Guangzhou science and technology project(201803010081).
E-mail addresses: isswy@mail.sysu.edu.cn(Yong Wang);
364085901@qq.com(Zhuoyi Su).

DenseNet with 135 layers. Experimental results show that it achieves better performance than those of the state-of-the-art efforts.

The remainder of this paper is organized as follows. The feature and the dense CNN of our proposed algorithm are presented in Section 2 and Section 3, respectively. Experiments are described in Section 4. The conclusions are given in Section 5.

2. MODEL OF VT SPOOFING

VT spoofing in many audio/speech editors is based on phase-vocoder methods [11] [12], in which a speech signal is represented by a quasi-stationary sinusoidal model computed from short-time Fourier transform (STFT). However, due to the resolution limitation, STFT bin frequencies generally do not represent true or instantaneous frequencies. Thus, a phase-vocoder is introduced which employs phase information that STFT ignores to improve frequency estimation, and to break the traditional tie between time and frequency characteristics to keep the tempo unchanged. VT spoofing can be depicted briefly as follows [11].

Suppose $x_t(n)$ is a frame of length N from the input speech signal at time t. Firstly, the FFT coefficients of $x_t(n)$ is obtained by Eq.1,

$$F(k) = \sum_{n=0}^{N-1} x_t(n) \cdot w(n) e^{-i\frac{2\pi kn}{N}} \quad 0 \leqslant k < N \tag{1}$$

where $w(n)$ denotes a Hamming or Hanning window and k denotes frequency bin index.

Then, instantaneous magnitude $|F(k)|$ and instantaneous frequency $\omega(k)$ are computed in Eq.2 and Eq.3, respectively,

$$|F(k)| = |\sum_{n=0}^{N-1} x_t(n) \cdot w(n) e^{-i\frac{2\pi kn}{N}}| \quad 0 \leqslant k < N \tag{2}$$

$$\omega(k) = (k + \Delta) \cdot F_s/N \quad 0 \leqslant k < N \tag{3}$$

where Δ denotes the deviation of the k^{th} bin frequency and F_s denotes the sampling frequency. The computation of Δ can be referred to in [12].

For VT spoofing, transient frequency $\omega(k)$ is modified by Eq.4, where α denotes the scale factor, i.e. the spoofing factor.

$$\omega'(\lfloor k \cdot \alpha \rfloor) = \omega(k) \cdot \alpha \quad 0 \leqslant k, k \cdot \alpha < N/2 \tag{4}$$

Linear interpolation is often used to modify the instantaneous magnitude, as seen in Eq.5 [12], where $0 \leqslant k, k' < N$ / 2, $k = [\ k'$ / $\alpha\ ]$, and $\mu = k'$ / $\alpha - k$.

$$|F'(\lfloor k \cdot \alpha \rfloor)| = \mu |F(k)| + (1 - \mu)|F(k + 1)| \tag{5}$$

Energy-preserving modification is another method to change the instantaneous magnitude by Eq.6.

$$|F'(\lfloor k \cdot \alpha \rfloor)| = \sum_{\lfloor k \cdot \alpha \rfloor \leqslant k \cdot \alpha < \lfloor k \cdot \alpha \rfloor + 1} |F(k)| \tag{6}$$

For simplicity, we still use k as the index of the modified instantaneous frequency ω' and the instantaneous magnitude F'.

Then the instantaneous phase $\phi'(k)$ is calculated via the instantaneous frequency $\omega'(k)$, and the transformed FFT coefficients are obtained by Eq.7.

$$F'(k) = |F'(k)| e^{i\phi'(k)} \tag{7}$$

Finally, the VT spoofed signal is obtained by inverse FFT performed on $F'(k)$.

From Eq.4 and Eq.5, we can see that VT spoofing modifies spectrum magnitude so that implicit features may be introduced into the spoofed speech signal. Therefore in our proposed method, we use the spectrogram of a speech as the input of a deep neural network to extract deep features for classification. We obtain the spectrogram of an input speech signal by STFT, where the window size is 175, and the overlapping is 50%. With respect to phonetics, VT spoofing is measured by a 12-semitone division [13] leading the spoofing factor α to the following form in Eq.8.

$$\alpha(s) = 2^{s/12} \tag{8}$$

s can take any integer value in the range of [-12, +12]. A modification too weak or too strong will result in deception failure or auditory unnaturalness. Therefore, in the experiments, we consider the medium ranges between [-8, -4] and between [+4, +8] that present the strongest deception ability.

3. DEEPLEARNING FRAMEWORK

3.1. The Dense Convolutional Network

In a conventional CNN, the output of the previous layer X_{l-1} is transmitted to the next layer as input by a non-linear operation H_l to get the output X_l.

$$X_l = H_l(X_{l-1}) \tag{9}$$

It is difficult to train a conventional CNN as degradation occurs with the increment of layers. To have a good inhibitory effect on the degradation, Residual Networks (ResNets) [14], Highway Networks [15] and FractalNets [16] create short paths X_{l-n} from early layers to later layers as shown in Eq.10.

$$X_l = H_l(X_{l-1}) + X_{l-n} \tag{10}$$

However, recent researches suggest that this type of connection leads to the fact that many layers contribute very little but occupy a large amount of computation [17]. Thus, an improved structure of ResNet named Dense Convolutional Network (DenseNet) was proposed to avoid this problem. In a DenseNet, any layer has direct connections to all subsequent layers, as shown in Eq.11,

$$X_l = H_l([X_0, X_1, ..., X_{l-1}]) \tag{11}$$

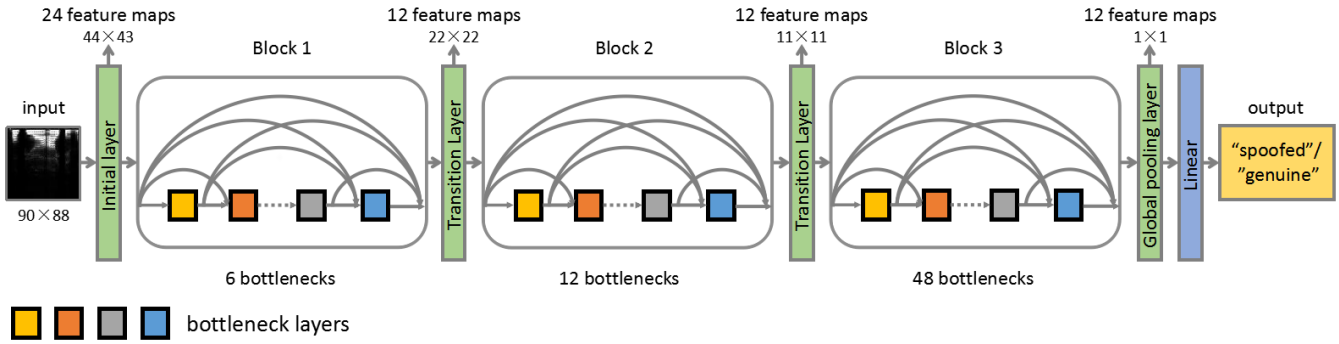

Fig. 1: The architecture of the proposed network

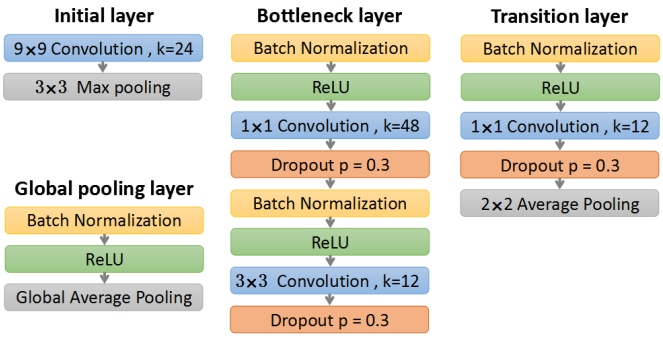

Fig. 2: The inner structures of each kind of the layers. Batch Normalization [18], ReLU, Dropout [19] and Pooling are of the operations before and after the convolution layer. k refers to the number of convolution kernels.

where X_0, X_1, X_{l-1} denote the output of the previous layer of layer l and $[...]$ denotes a the concatenation operation. Furthermore, the output dimension of each layer has k feature maps, where k is usually set to a small value.

This kind of dense connection mode have significant advantages over the aforementioned networks: 1) It ensures maximum information flow between layers to strengthen feature propagation. 2) Dense connections have a regularizing effect which reduces over-fitting on tasks with smaller training set sizes. 3) It allows DenseNet layers to be narrow, e.g. $k = 12$, to significantly reduces the number of parameters, to alleviates the problem of degradation, and to support the reuse of limited neurons. 4) It does not need to relearn redundant feature-maps and is convenient for training.

3.2. Structure of the Proposed DenseNet

The structure of the proposed DenseNet is shown in Fig.1. The inputs are single channel spectrogram obtained by STFT, and the sizes are all set to 90×88. The network consists of an initial layer, three dense blocks, two transition layers, a global pooling layer and a linear layer. The three dense blocks consist of 6, 12 and 48 bottleneck layers, respectively. The

Table 1: The number of 1s clips in each set

Dataset	Clip number	Dataset	Clip number
TIMIT_1	7996	TIMIT_2	8967
NIST_1	18601	NIST_2	14589
UME_1	7482	UME_2	6952

linear layer is a full connection layer followed by a softmax with two outputs which represent the probabilities of 'genuine' and 'spoofed', respectively. The inner structures of each kind of these layers are shown in Fig.2. Each bottleneck layer contains 2 convolution layers, so that the entire DenseNet contains 2×(6+12+48) + 1 + 1 + 1 = 135 convolution layers.

A bottleneck layer contains a 1×1 convolution layer followed by a 3×3 convolution layer instead of two 3×3 convolution layers to reduce computation, as shown in Fig.2. The transition layer connects two adjacent denseblocks to further reduce the size of the feature-maps.

4. EXPERIMENTAL RESULTS

4.1. Experimental Corpora and Setup

Three corpora are used in the experiments, namely, Timit(6,300 clips, 630 speakers), NIST(3560 clips, 356 speakers) and UME(4040 clips, 202 speakers). They are of WAV format, 8 kHz sampling rate, 16-bit quantization and mono. Each corpus is divided into 2 sets as follows.

Training set: Timit-1 (3000 clips), NIST-1(2000 clips), UME-1(2040 clips);

Testing set: Timit-2(3300 clips), NIST-2(1560 clips), UME-2(2000 clips).

Each clip is further cut into several 1s clips. The number of 1s clips in each set is shown in Table 1.

Four prevailing spoofing methods, i.e., Audacity [20], Cool Edit [21], PRAAT [22] and RTISI [23], with spoofing factors between [-8, -4] and between [+4, +8] are taken into consideration. Thus, the number of the spoofed (negative) clips is 40 times that of genuine (positive) clips. In order to achieve data balance, we expand the positive clips by shifting

Table 2: The detection accuracy of intra-database evaluation

Training dataset	Testing dataset	The proposed method	Liang's method [25]	Wu's method [26]
TIMIT_1	TIMIT_2	99.45%	96.52%	95.87%
NIST_1	NIST_2	98.04%	95.93%	94.56%
UME_1	UME_2	97.56%	94.85%	93.63%
Average	Average	98.35%	95.77%	94.69%

every other 200 samples to have the same number of positive clips as negative clips.

We use an ADAM optimizer [24] to train the proposed DenseNet with L_2 loss function, β_1 and β_2, namely, the exponential decay rates for the 1st and the 2nd moment estimates, are set to be 0.9 and 0.999, respectively. The epsilon hat ϵ is set to be 10^{-8}. The learning rate and the dropout rate are set to be 10^{-4} and 0.3, respectively. The training batches are 100,000 and the batch size is 64.

Detection accuracy in Eq.12 is used to measure the performance,

$$d = (G_d + S_d)/(G + S) \tag{12}$$

where G and S are the numbers of genuine and spoofed clips in the testing sets, respectively, and G_d and S_d are the numbers of the genuine clips correctly detected from G and of the spoofed clips correctly detected from S.

4.2. Intra-database evaluation

In the case of intra-database, the testing set and training set are from the same corpus. The detection results of our proposed method and the other reported efforts are shown in Table.2, from which we can be seen that the average detection accuracy of ours is 2.58% higher than that of [25] which adopts conventional CNN model, and 3.66% higher than that of [26] which employs SVM.

Our proposed method outperforms the other two, because a DenseNet model has much more layers than a conventional CNN so that it can extract more and deeper features to facilitate classification. Besides, in a conventional CNN, decision is made with deep features solely. But in a DenseNet, due to the dense connection mode, the decision is made with deep features as well as early edge features so that accuracy can be further improved.

4.3. Cross-database evaluation

In reality scenarios, testing speech and training speech may come from different sources, and thus they may have different intrinsical features. Therefore, cross-database evaluation is conducted to test the diversity of the proposed method. One of the 3 corpora is selected as the testing data set and the other two are as the training sets. The experimental results are shown in Table 3. We can see that the results of the first two

Table 3: The detection accuracy of cross-database evaluation

Case	Training dataset	Testing dataset	The proposed method
Case 1	TIMIT_1/NIST_1	UME_2	96.45%
Case 2	NIST_1/UME_1	TIMIT_2	95.26%
Case 3	TIMIT_1/UME_1	NIST_2	80.20%
	Average	Average	90.63%

cases are quite good, but case 3 is not ideal. One possible reason is that the data volume of NIST is larger than the other two sets as shown in Table 1, and the model trained by NIST has better generalization ability. In [25], the accuracy of case 1 is given as 94.37%, while ours is 96.45% indicating that our proposed method outperforms the one in [25]. The results of the other two cases are not presented in [25].

5. CONCLUSION

In this paper, a method based on dense convolutional network for detecting spoofed speech from genuine speech is presented. Deep features can be automatically extracted by the 135-layer DenseNet. The experimental results indicates that it is superior to the state-of-the-art methods, and it achieves computing efficiency by careful optimization of kernel reduction and by the employment of bottleneck layers. The future work will focus on the application of deeper network structure to extract deeper features and further improve the accuracy.

6. REFERENCES

[1] Patrick Perrot, Guido Aversano, and Grard Chollet, "Voice disguise and automatic detection: Review and perspectives," in *The Workshop on Progress in Nonlinear Speech Processing*, 2007, pp. 101–117.

[2] Hannah Muckenhirn, Mathew Magimai-Doss, and Sebastien Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–6.

[3] Hannah Muckenhirn, Pavel Korshunov, Mathew Magimai-Doss, and Sbastien Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 25, no. 11, pp. 2098–2111, 2017.

[4] Jess Villalba and Eduardo Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE International Carnahan Conference on Security Technology*, 2011, pp. 1–8.

[5] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sbastien Marcel, "Joint speaker verification and antispoofing in the i -vector space," *IEEE*

Transactions on Information Forensics & Security, vol. 10, no. 4, pp. 821–832, 2015.

[6] Dipjyoti Paul, Monisankha Pal, and Goutam Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 605–617, 2017.

[7] Zhizheng Wu, Phillip De Leon, Cenk Demiroglu, Ali Khodabakhsh, Simon King, Zhen Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, and Mirjam Wester, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[8] Cenk Demiroglu, Osman Buyuk, Ali Khodabakhsh, and Ranniery Maia, "Post-processing synthetic speech with a complex cepstrum vocoder for spoofing phase-based synthetic speech detectors," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[9] Chunlei Zhang, Chengzhu Yu, and John H. L Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2017.

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

[11] Yong Wang, Haojun Wu, and Jiwu Huang, *Verification of hidden speaker behind transformation disguised voices*, Academic Press, Inc., 2015.

[12] Jean Laroche, "Time and pitch scale modification of audio signals(chapter 7)," in *International Series in Engineering and Computer Science*. 2006, vol. 437, pp. 279–309, Springer US.

[13] Sandra E Trehub, Annabel J Cohen, Leigh A Thorpe, and Barbara A Morrongiello, "Development of the perception of musical relations: Semitone and diatonic structure.," *Journal of Experimental Psychology Human Perception & Performance*, vol. 12, no. 3, pp. 295, 1986.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Training very deep networks," *CoRR*, vol. abs/1507.06228, 2015.

[16] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *CoRR*, vol. abs/1605.07648, 2016.

[17] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger, "Deep networks with stochastic depth," *CoRR*, vol. abs/1603.09382, 2016.

[18] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," pp. 448–456, 2015.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] "Audacity:free audio editor and recorder," `http://audacity.sourceforge.net`, 2012.

[21] "Cool edit pro is now adobe audition," `http://www.adobe.com/products/audition.html`, 2012.

[22] "Praat: Doing phonetics by computer," `http://www.fon.hum.uva.nl/praat`, 2012.

[23] "Time-scale / pitch modification tools," `http://www.mathworks.com/matlabcentral/fileexchange/25880-time-scalepi`, 2012.

[24] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[25] Huixin Liang, Xiaodan Lin, Qiong Zhang, and Xiangui Kang, "Recognition of spoofed voice using convolutional neural networks," in *IEEE Global Conference on Signal and Information Processing*, 2017, pp. 293–297.

[26] Haojun Wu, Yong Wang, and Jiwu Huang, "Blind detection of electronic disguised voice," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3013–3017.