AN ATTEMPT OF CNN ON SPEAKER VERIFICATION

TIAN Yi, ZHUO Qing

Department of Automation, Tsinghua University, Beijing, 100084, China
tianyi15@mails.tsinghua.edu.cn; zhuoqing@tsinghua.edu.cn

ABSTRACT

In this paper, we examine the ability of convolutional neural networks (CNNs) to extract human voice features, and use these deep features as well as fully connected networks to tackle the problem of speaker verification. We demonstrate that CNNs are promising for the extraction of voice features, which can be of great importance in further researches and applications.

***Index Terms*—** Speaker verification, deep learning, convolutional neural networks

1. INTRODUCTION

Speaker verification has long been an exciting research problem, considered part of the problem of speaker recognition, whose goal is to extract, characterize and recognize the information about speaker identity in the speech signal [1]. Different than speaker identification, which is typically a close-set problem, whose aim is to identify the speaker among a set of known speakers, the problem of verification is typically open-set, aiming to determine whether or not a person is who s/he claims to be.

This has been an active problem for over five decades, and the most prevalent application of this technique is to use it as a biometric to control the access in certain circumstances. [1] What stimulates us to research in this problem is, however, this control ability in more loving and romantic situations, such as a gift box from a boy to a girl which can open only when the girl — not any other girl — speaks.

Convolutional neural networks (CNNs), considered an effective way to extract features in image processing, have been used for a variety of other purposes, recently for machine translation. In this paper, we aims to make a tentative research into its ability in the problem of speaker verification. To investigate this problem, we first collect a dataset made up of audios from 3 subjects, and then set up several CNNs to conduct experiments. Our result shows that CNNs are promising in this task and achieved an accuracy rate of 100% on train and validation data and the same 100% on test data in our proposed dataset. It should be noted that we collect data from our friends due to our attempt to make it more close to our daily life. And limited to the simple purpose of this research, we make no comparison to other publicly available datasets. Our codes have been made available on `http://github.com/siarnold/SS`.

2. DATASET

To examine the ability of convolutional networks on the problem of speaker verification, we propose a dataset including 10 audios, 75,955,008 frames and 2,207 seconds in the aggregation. Despite the various sources of the data, the sampling rates ranging from 8,000Hz to 48,000Hz, the number of channels being 1 or 2, such a diversity can be uniformed with 'ffmpeg' transformation. The 3 subjects include two females with 4 and 3 audios respectively, and one male with 3 audios. Granted, the dataset is limited in terms of quantity or variety. However, as part of a job aiming to make a simple research into the ability of CNN on this task, such a dataset would be considered moderately acceptable. It should also be noted that every subjects' records include more than one kind of the languages in the range of Chinese, English and German, which makes this dataset more fascinating and challenging.

Apart from the dataset used for the training and testing of the CNNs, we also record an audio to test their abilities on audios recorded in other situations. In the experiments introduced below, we select one of the subjects to be the person to be verified.

3. MODEL FORMULATION

According to the course on Sensory Physiology of Indiana University [2], there are usually 3 orders of neurons in the conduction pathway of an auditory signal to CNS. However, to handle the speaker verification task, the number of orders of related neurons in CNS is of real importance, which is so complex a problem that to our knowledge biological studies have not yet provided a clear answer. The same problem in biological understandings is confronted in computer vision, without hindering the progress of its neural network modeling. Thus, we hereby propose a neural network model based not on the biological structure, but on our assumption of the auditory understanding process. We propose 3 models as follows.

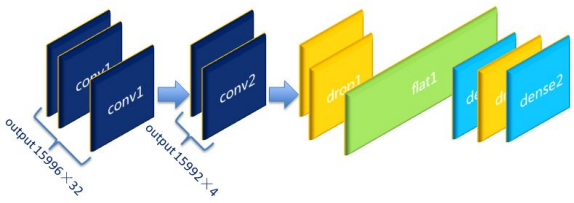

Fig. 1. **The first proposed model.** The model is composed of two blocks, the first half being convolutional and the second half being fully or densely connected.

spv_cnn_v0 As shown in Fig. 1, spv_cnn_v0 consist of two blocks. The first two 1-dimensional convolutional networks with kernel sizes of 5×32 and 5×4 respectively without zero-padding and a drop-out layer, designed for solving the problem of overfitting , form the first block, which intends to extract deep features using convolutional computation. Before passing the features to a fully-connected network, we flatten the the 15992×4 features to one dimension with a transitional layer. Then comes the second block which is comprised of one fully-connected layer of 128 neurons, one drop-out layer and another fully-connected layer of 2 neurons, which is a classical perceptron model. The design of such a network is inspired by the mnist CNN model [3], with the consideration of the complexity of the speaker verification problem. This model has 8,189,126 trainable paramters in total.

spv_cnn_v1 The second model shares the same first block as spv_cnn_v0, but is designed with greater complexity for the second block, which is comprised of three fully connected layers of size 1024, 1024 and 2 respectively. Between the adjacent fully connected layers is added one drop-out layer. In total the spv_cnn_v1 model has 66,556,742 trainable parameters, which is much larger a number than the previous network.

spv_cnn_vf The third model is the most simple model of the three, with minor changes of spv_cnn_v0. The two convolutional layers have both a kernel size of 3×3, while the two fully connected layers have the size of 64 and 2 respectively.

We design the three CNNs above to investigate the ability of networks of different complexities on this problem. The most simple network (spv_cnn_vf) is proven by experiment to be effective enough for this task, as will be discussed below.

4. EXPERIMENTS

4.1. Data Preprocessing

Before we use the audios to train or test the models, we first segment the audios into 4-second audio segments, so that the frequency resolution can reach 0.25Hz. From the 10 audios we derive 548 audio segments, which are then uniformly sampled at a frame rate of 8,000 Hz before being preprocessed with FFT (Fast Fourier Transform). In this way, we acquire an upper distinguishable frequency of 4,000 Hz by Nyquist-Shannon Sampling Theorem. Normally, human voices have a fundamental frequency of 85 to 255 Hz. However, to distinguish between different people by their voices, more dedicate differences like high-order harmonic series need to be considered, which can range between 300 Hz and 3,000 Hz. [4] For this reason, we consider our upper frequency selection to be necessary and practical. Actually, we find that it is possible for humans to distinguish between people by listening to the 4,000-Hz sampled audios. As is known, FFT is an effective method to transform the audio wave from time-domain to frequency-domain, which we use to make the features more salient. Since the amplitudes of FFT, which we use, are evenly symmetric, meaning that the second half of the transformed data are the same as the first half, we obtain the input of a 16,000-dimensional vector.

In order to evaluate the effectiveness of this method, we divide the 548 segments into three lists: the train list, the validation list and the test list. While the train list and validation list are used in the training, the test list are specially used after training to evaluate whether the model can really solve this verification problem. To escape the problem of using segments from the same audio to train and test, we select a Chinese, an English and a German audio from the three subjects respectively for testing, others being randomly distributed at a proportion of 3:1 for training and validating.

4.2. Training

We use an nvidia GeForce GTX 1080 Ti GPU to train this network using keras with tensorflow backend. The details of our training parameters include: stochastic gradient descent optimizer with learing rate of 1e-6, decay of 3e-9, momentum of 0.9, with nesterov, and the classical cross-entropy loss function. Using 90% of the resources of the GPU, the training costs approximately 7 minutes.

The three models were trained for 300 epoches, as shown in Fig. 2. The spv_cnn_v0 model and spv_cnn_vf model showed desirable results and reached the accuracy of 1.00 in the end. Due to the uncertainty in the drop-out layers, each training process for the same model may differ from each other, still possibly reaching the same converging point. However, the most complex network spv_cnn_v1, instead of converging to a global optimal, only reached a local optimal, which should be attributed to its huge quantity of parameters and inadequate data. We tried to escape the local optimal by increasing the learing rate (as well as its decay) to 1e-5 and 1e-4, which only showed a slightly better or even worse performance, converging to an accuracy rate of 70.70% and 29.3% respectively. Based on this, we lower our estimation of the complexity of this task and design the network of spv_cnn_vf.

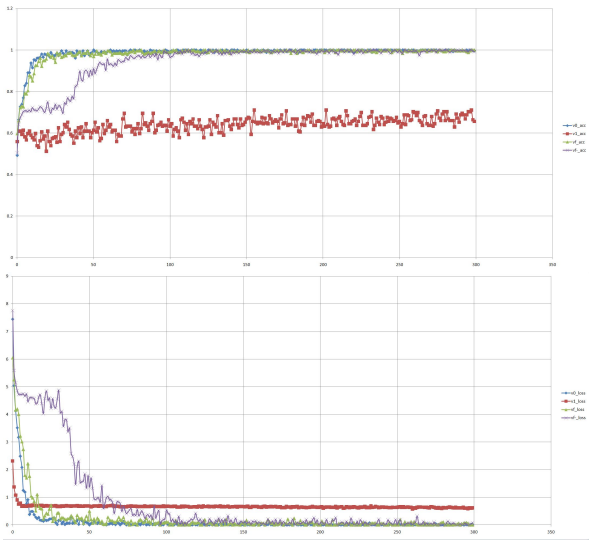

Fig. 2. **The accuracies and losses of the models during training on the data in the train list.** The figure includes the training processes of spv_cnn_v0, spv_cnn_v1 and spv_cnn_vf. The model spv_cnn_vf was trained twice denoted by vf and vf- respectively and showed different converging processes, with the same converging point nonetheless.

Table 1. **The accuracy rates** on the test data of spv_cnn_v0 model and spv_cnn_vf model.

epoch	50	100	200	300
spv_cnn_v0	100.00%	100.00%	100.00%	100.00%
spv_cnn_vf	100.00%	100.00%	100.00%	100.00%

4.3. Results

We test the two trained models of spv_cnn_v0 and spv_cnn_vf on the test data generated in the way mentioned in section 4.1. The results are shown in Table. 1. Both the two models achieve an accuracy rate of 100% after 300 epochs of training, which is a delightful success considering that the three test audios are in Chinese, English and German respectively, not essentially the same languages used to train the models. This further demonstrates that our method extracts the essential features of voice timbres.

As mentioned in section 2, we also use another audio to test the models' abilities on audios recorded in other situations. To our disappointment, this audio, recorded by the person to be verified, was not successfully verified by our models. The reasons could include the variance of voices (making this task hard even for humans), the varied recorded conditions (inadequate dataset potentially causing overfitting for this task) and lack of pre-training (not the same process as we humans learn to distinguish voices in which we first learn to distinguish voices from other sounds).

5. CONCLUSION AND FUTURE WORK

In this paper, we make a tentative research on the problem of speaker verification. We first collect a dataset with our classmates, and then set up three CNN models to experiment. From the experiments, we can see that CNNs are a powerful and effective model to extract voice features. Combined with fully connected networks, these features are useful for the task of speaker verification.

In the future, to further address this problem, larger datasets need to be utilized and filtering methods need to be applied. The networks would hopefully perform better and have less of a problem of reaching the local optimal if the networks are pre-trained for human voice detection, as the previous experience for this task.

Acknowledgements Special thanks should be given to Qu Yinsheng and Li Chendi, who contribute a lot in the collection of the dataset.

6. REFERENCES

[1] D. A Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. IV–4072–IV–4075.

[2] "Sensory physiology, chapter 10," `http://www.indiana.edu/~nimsmsf/P215/p215notes/PPlectures/Printables/SENSES.pdf`.

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] "Voice frequency," `https://en.wikipedia.org/wiki/Voice_frequency`.