D-vector based speaker verification system using Raw Waveform CNN

Jeeweon Jung [1, a], Heesoo Heo [1, b], Ilho Yang [1, c],
Sunghyun Yoon [1, d], Hyejin Shim [1, e], and Hajin Yu [1, f]

[1] Department of Computer Science, University of Seoul, South Korea

[a]jeewon.leo.jung@gmail.com, [b]zhasgone@naver.com, [c]heisco@daum.net,
[d]ysh901108@naver.com, [e]shimhyejin930615@gmail.com, [f]hjyu@uos.ac.kr

Keywords: d-vector, speaker verification, raw-audio-CNN

Abstract. In this paper, we propose a d-vector based speaker verification system in which raw-audio-CNN is used as a d-vector extractor instead of a conventional multi-layer perceptron. Because raw-audio-CNN takes raw wave signals as input, traditional acoustic feature extracting methods such as mel-frequency cepstral coefficient and mel-filterbank features are no longer needed. The results of experiments conducted show that raw-audio-CNN can successfully perform functions carried out by traditional acoustic feature extracting methods and outperforms traditional d-vector systems that utilize standard multi-layer perceptron with acoustic features.

1 Introduction

Speaker recognition is a field in which the characteristics of an utterance are used to identify speakers. Speaker recognition is usually divided into two sub-fields: speaker identification and speaker verification. Speaker identification is an *n*-classification task in which *n* speakers exist and the goal is to determine the speaker of an input utterance. Speaker verification is a binary classification task that verifies whether a claimed speaker and a proposed test utterance have the same identity.

As in many other fields, adoption of deep neural networks (DNNs) is being extensively researched in speaker verification [1][2]. Among these approaches using DNN, Variani et. el. recently proposed a d-vector based speaker verification system in which a speaker identifier DNN is first trained with a development set to identify speakers. After training the speaker identifier DNN, it is used to extract speaker identity representation, d-vector, of unseen speakers from test sets [3].

The d-vector based system has become one of the most popular approaches in the adoption of DNN to speaker verification and various researchers have extended the d-vector system. The various extensions can be classified into two main types: (1) Enhancement of speaker identifier DNN and (2) replacement of subtasks previously carried out using conventional methods with DNN. Unlike the d-vector system, which use a multi-layer perceptron (MLP) as a speaker identifier DNN, systems using convolutional neural network (CNN) or recurrent neural network (RNN) have been explored with successful results [4][5]. As regards the latter type of extension, conventionally handcrafted acoustic features such as mel-frequency cepstral coefficient (MFCC) and mel-filterbank are being used to train speaker identifier DNN and to extract d-vectors in d-vector based speaker verification systems. Alternatively, in some systems, spectrograms and even raw audio are being used as input to the neural network for speech recognition [6] and music tagging [7].

This paper expands on the above systems and proposes a d-vector based speaker verification system in which raw waveform signals are used as input to the speaker identifier DNN. The concept of multi-task learning is applied throughout this paper following study results that indicate that multi-task learning enables speaker identifier DNN to extract more general and robust speaker identity representation [8][9].

The remainder of this paper is organized into the following sections. The related work section introduces existing studies that use raw audio as the feature in training DNNs. The d-vector based speaker verification system section addresses the overall flow of d-vector systems. The multi-task learning in speaker verification section explains the basic concept of multi-task learning and why it works for d-vector based systems. The raw audio convolutional neural network section reveals the architecture of the speaker identifier CNN that takes raw audio wave samples as input. The experimental settings and results section deals with training configuration, experimental settings, and results. Finally, conclusion and future work are presented.

2 Related work

Systems using raw audio as input to DNN have been proposed in various domains, such as speech recognition[6], music classification, and audio tagging [7]. For example, raw audio has been exploited as features in speech recognition [6] and also for music auto-tagging [7]. More specifically, the concept of a 1D strided convolution layer that specially designs the first convolution layer for raw audio signals has been proposed [7]. Unlike conventional convolution layers, a 1D strided convolution layer takes only 3–6 samples that account for 0.18–0.36 ms for audios with a sampling rate of 16,000 Hz. This sample-level approach is intended to imitate other convolution models in the image or text domain where input is given as pixels at the character level. This sample-level approach in raw audio waveforms has been proven effective and is thus used in this research.

2.1 D-vector based speaker verification system

Conventional d-vector based speaker verification consists of the following steps. First, pre-processing is conducted by extracting acoustic features and applying voice activity detection (VAD) and normalization techniques such as mean and variance normalization (MVN). Following preprocessing, a speaker identifier DNN is trained using the utterances from a development set. It is worth noting that when the speaker identifier DNN is trained, the input acoustic feature is concatenated with context frames, meaning that a few of the previous frames and the ensuing frames are also placed into the speaker identifier DNN. Context frames are concatenated because a single frame, which is typically accounts for 25 ms, does not have sufficient information about the speaker.

After the speaker identifier DNN is trained, the output layer is removed. Then, enrollment and test utterances are forward-propagated through the speaker identifier DNN to the last hidden layer. After forward propagation, the activations of the last hidden layer is used as d-vector for the given input. However, this d-vector is extracted from a few frames (1 + number of context frames) and is therefore considered as being frame-level. Only one d-vector is extracted from one utterance—obtained by element-wise averaging frame-level d-vectors extracted from the same utterance.

Speaker models are then composed by element-wise averaging the enrollment utterances' d-vectors for a given speaker. Finally, cosine similarity scoring is applied between a speaker model and a given test d-vector and similarity is measured. The overall process used in the d-vector based speaker verification system is illustrated in Figure 1.

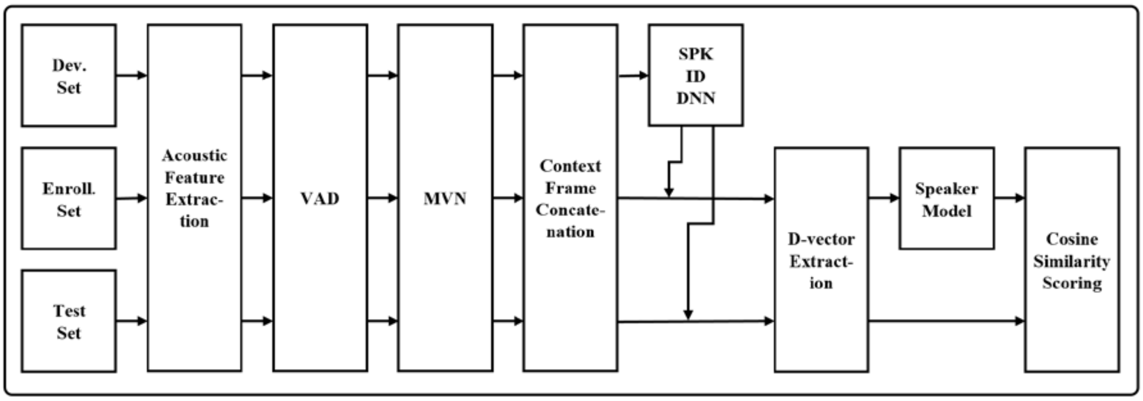
Dev. Set
Enroll. Set
Test Set
Acoustic Feature Extrac-tion
VAD
MVN
Context Frame Concate-nation
SPK ID DNN
D-vector Extract-ion
Speaker Model
Cosine Similarity Scoring

Fig. 1. Pipeline process employed in conventional d-vector based speaker verification

2.2 Multi-task learning in d-vector based speaker verification (baseline)

In multi-task learning, several tasks are jointly optimized by utilizing more than one output layer. The core assumption in this joint optimization is that tasks that are optimized together are relevant; thus, jointly training them can produce synergy. This synergy can also be developed in speaker verification. By adopting multi-task learning, more information, typically phrase or phoneme information, is additionally provided to the hidden layers in the training phrase. For example, in the speaker identifier DNN, consider a case in which both speaker and phrase are being jointly classified. The network will be optimized in a manner that holds features to classify both speaker and phrase. During the progress, information to classify phrases will be helpful for classifying speakers and vice versa.

Synergy coming from multi-task learning can be verified by comparing the performance of d-vectors extracted from a single task DNN and a multi-task DNN. In [8], experimental results show that adopting multi-task learning enhances d-vector based speaker verification systems. Therefore, it can be said that multi-task learning provides an opening that allows humans to give additional information to neural networks when the additional information is considered helpful.

2.3 Raw audio convolutional neural network in d-vector based speaker verification (proposed)

The pipeline process employed by the proposed d-vector based speaker verification system is depicted in Figure 2. Compared with Figure 1, it can be seen that the acoustic feature extraction, VAD, MVN, and context-frame concatenation processes have all been removed.

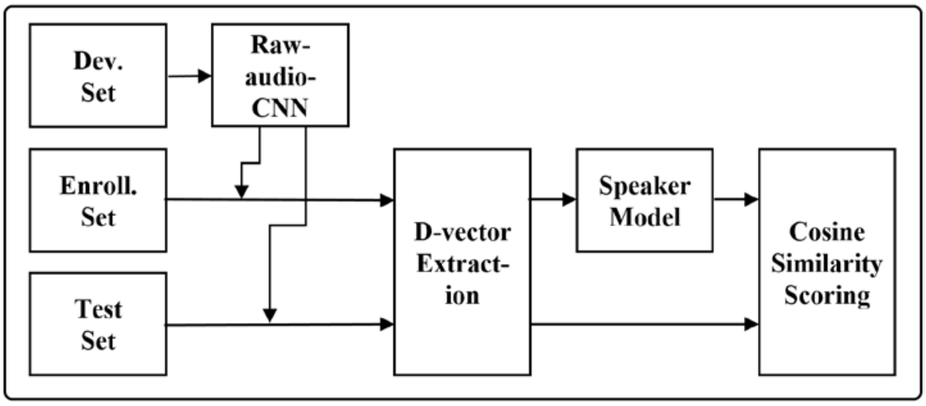

Fig. 2. Pipeline process employed by the proposed d-vector based speaker verification using raw-audio-CNN system

The speaker identifier DNN architecture exploited in this paper primarily comprises convolution layers. Strided convolution, proposed in [7], is used as the first hidden layer to process raw audio signals. Strided convolution has a short filter size of three, and the stride is also three. This means that no overlapping exists and the output length of this layer will be one-third of the input data. This approach is meant to consider three samples accounting the same with one pixel in each image. In addition, the pooling layer is not attached after the strided convolution layer.

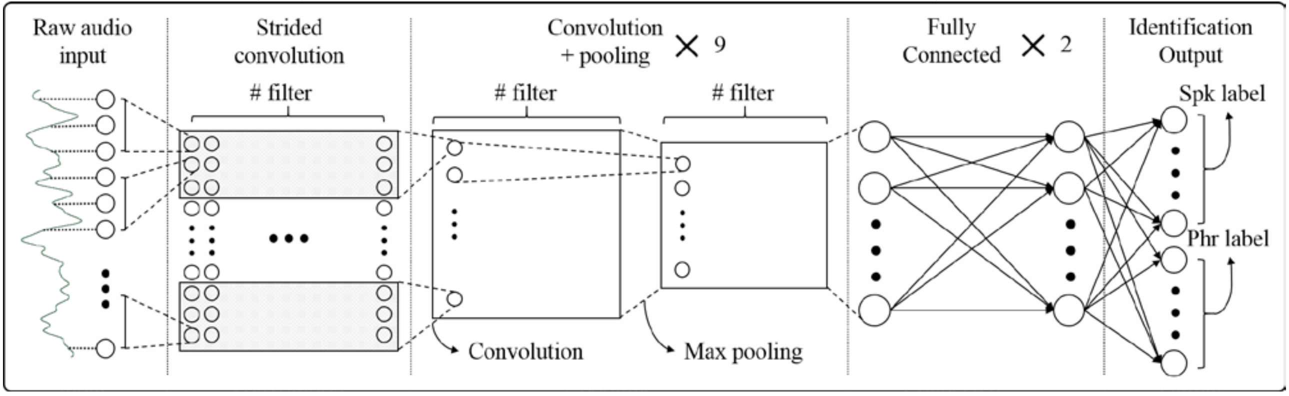

Fig. 3. Overall process flow of exploited raw-audio-CNN

After the strided convolution layer, conventional convolution layers with pooling layer attached are exploited. These convolution layers have a short filter of size three and a stride of one that does not reduce the length of the input. Instead, the max pooling layer reduces the length of the data. This approach of repetitively using convolution layers with small filters is based on the same hypothesis as [7], in that short filters of the same small size of three can be efficient for the same reason as VGG [9] net in that deep CNNs will learn hierarchical representations using its depth. After the convolution layers, two fully connected layers are added before the output layer. The utilized network is illustrated in Figure 3.

As the number of input nodes cannot be changed, raw audios that are longer than the input length are truncated and, if the length of the raw audio is shorter than that of the CNN's input, it is duplicated and truncated to match the length of the input layer. End point detection (EPD) is also utilized to cut off silence sections at the start and end of the audio. Note that because audio signals are 1D, the data are treated as 2D data with a height of one.

3 Experimental settings and results

Theano [10, 11], a deep learning library in python, was used in all the experiments presented in this paper. Part 1 of RSR2015 [12], a widely used dataset for research on speaker verification, was selected as the database in this study. Part 1 of RSR2015 comprises 300 speakers: 157 males and 143 females. A total of 270 utterances, composed of thirty different short sentences, each on average about 3 seconds, are provided for each speaker, using different channels. Speaker model composition followed the 3sesspwd-eval of RSR2015 and trial selection followed 3sess-pwd script of RSR2015 guidelines.

The d-vector extracted from standard MLP with multi-task learning of speaker and phrase classification was used as the baseline. The baseline network comprised seven hidden layers, each having 1024 nodes. L2 weight decay [13], Dropout [14], and learning rate decay were applied. In the baseline, a 48-dimensional mel-filterbank was extracted as the acoustic feature and VAD and MVN were applied sequentially. Thirty-five previous frames and 12 ensuing frames were used as context frames during training of the speaker identifier DNN and extraction of the d-vectors.

The structure and operation of the proposed raw-audio-CNN based d-vector system are as follows. The strided convolution layer processes raw audio input at the first hidden layer. This layer has a stride and filter size of three, and no pooling layer is attached. This is followed by nine 1D convolution layers of stride one with a max pooling layer of stride three. All convolution layers have a filter of size three. Two fully connected layers with a dropout of 0.5 are attached next, before the output layer. Multi-task learning of speaker and phrase classification is also applied. Batch normalization is applied in all convolution layers. The details are the same as described in [7], except that the speaker identifier network in this paper has two fully connected layers. Experimental results are listed in Table 1. "d-vector (baseline)" is the d-vector system that used a multi-task based speaker identifier DNN and "d-vector (proposed)" is the proposed d-vector system using raw-audio-CNN. Equal error rate (ERR) is a widely used performance measure in speaker verification, where lower EER indicates good performance. The table shows that the d-vectors extracted from raw-audio-CNNs have a 7.61% EER compared to 8.34% EER for the d-vectors extracted from the baseline system.

Table 1. Equal error rate (ERR) of the baseline and the proposed system

System	EER (%)
d-vector (baseline)	8.34
d-vector (proposed)	**7.61**

4 Conclusion and future work

In this paper, d-vectors extracted from sample-level raw-audio-CNN were explored with the objective of replacing acoustic feature extraction and preprocessing with raw-audio-CNN. Using raw-audio-CNN, EER was reduced from 8.34% to 7.61%, as shown in the results of experiments conducted with short sentences in part 1 of the RSR2015 dataset.

Our future work will involve extending the backend to compose a complete end-to-end network in which a few utterances' audios are provided as input and the verification result is directly produced.

Acknowledgement

This research was supported by Projects for Research and Development of Police science and Technology under Center for Research and Development of Police science and Technology and Korean National Police Agency funded by the Ministry of Science, ICT and Future Planning (Grant No. PA-J000001-2017-101)

References

[1] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), pp.1527-1554.

[2] Seide, F., Li, G., Chen, X., & Yu, D. (2011, December). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on pp. 24-29. IEEE.

[3] Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014, May). Deep neural networks for small footprint text-dependent speaker verification. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on pp. 4052-4056. IEEE.

[4] Chen, Y. H., Lopez-Moreno, I., Sainath, T. N., Visontai, M., Alvarez, R., & Parada, C. (2015). Locally-connected and convolutional neural networks for small footprint speaker recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

[5] Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016, March). End-to-end text-dependent speaker verification. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on pp. 5115-5119. IEEE.

[6] Palaz, D., Doss, M. M., & Collobert, R. (2015, April). Convolutional neural networks-based continuous speech recognition using raw speech signal. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on pp. 4295-4299. IEEE.

[7] Lee, J., Park, J., Kim, K. L., & Nam, J. (2017). Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms. arXiv preprint arXiv:1703.01789.

[8] Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. (2015). Deep feature for text-dependent speaker verification. Speech Communication, 73, pp. 1-13.

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[10] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A. & Bengio, Y. (2012). Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590.

[11] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G. & Bengio, Y. (2010, June). Theano: A CPU and GPU math compiler in Python. In Proc. 9th Python in Science Conf pp. 1-7.

[12] Larcher, A., Lee, K. A., Ma, B., & Li, H. (2012). RSR2015: Database for text-dependent speaker verification using multiple pass-phrases. In Thirteenth Annual Conference of the International Speech Communication Association.

[13] Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In

Advances in neural information processing systems pp. 950-957.

[14] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), pp. 1929-1958.