A Health-friendly Speaker Verification System Supporting Mask Wearing

Chaotao Chen,[1] **Di Jiang,** [1] **Jinhua Peng,** [1] **Rongzhong Lian,** [1] **Chen Jason Zhang,** [2] **Qian Xu,** [1] **Lixin Fan,** [1] **Qiang Yang** [2]

[1] AI Group, WeBank Co., Ltd, Shenzhen, China
[2] Department of CSE, The Hong Kong University of Science and Technology, Hong Kong, China
{chaotaochen, dijiang, kinvapeng, ronlian, qianxu, lixinfan}@webank.com, jasonzhang@ust.hk, qyang@cse.ust.hk

Abstract

We demonstrate a health-friendly speaker verification system for voice-based identity verification on mobile devices. The system is built upon a speech processing module, a ResNet-based local acoustic feature extractor and a multi-head attention-based embedding layer, and is optimized under an additive margin softmax loss for discriminative speaker verification. It is shown that the system achieves superior performance no matter whether there is mask wearing or not. This characteristic is important for speaker verification services operating in regions affected by the raging coronavirus pneumonia. With this demonstration[1], the audience will have an in-depth experience of how the accuracy of bio-metric verification and the personal health are simultaneously ensured. We wish that this demonstration would boost the development of next-generation bio-metric verification technologies.

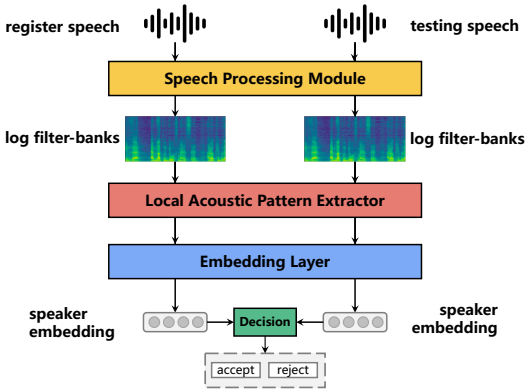

Figure 1: Architecture of the speaker verification system.

Introduction

Speaker Verification (SV) (Hansen and Hasan 2015) technologies aim to confirm a claimed speaker by analyzing his or her speech. It has been used in a wide range of real-world applications such as call center, risk management, mobile payment and smart device activation, etc.

In the face of contagious disease such as the coronavirus pneumonia, traditional bio-metric authentication techniques such as face verification put the users' heath at risk since masking wearing is not typically supported. In this demonstration, we propose a speaker verification system, which is robust in complex scenarios where the speaker's mouth and nose might be affected by the masking wearing. Specifically, the system is composed of a speech processing module, a local acoustic feature extractor and an embedding layer. In order for discriminative speaker embedding, the system leverages a powerful ResNet (He et al. 2016) neural network architecture as the feature extractor to models the local short spans of acoustic features. In particular, a novel multi-head attention embedding layer is introduced to integrate the speaker-specific local patterns into speaker embedding. The proposed embedding turns out to be more effective than those based on statistical pooling (Snyder et al. 2018) or attentive pooling (Okabe, Koshinaka, and Shinoda 2018). Moreover, our system is optimized by an additive margin softmax loss (Wang et al. 2018a) to discriminate between speakers. Thus, it can capture more discriminative and robust speaker embedding than those general DNN-based SV methods using cross entropy loss (Nagrani, Chung, and Zisserman 2017), triplet loss (Novoselov et al. 2018) or center loss (Li et al. 2018). The demonstration shows that the system is not only superior in traditional speaker verification scenarios but also robust in complex scenarios such as mask wearing.

System Architecture

The system architecture is illustrated in Figure 1. The system mainly consists of three major components: (1) a speech processing module that transforms the raw speech into acoustic features (e.g. filter-bank and MFCC); (2) a local acoustic pattern extractor that takes variable-length acoustic features as input and encodes them as a sequence of frame-level representations with respect to local patterns; (3) an embedding layer that aggregates the frame-level features into a fixed-dimensional utterance-level speaker embedding.

Specifically, given raw speech acquired from the microphone in WAV format, the speech processing module first employs Voice Activity Detection (VAD) to filter out non-speech frames from the speech input, and then transforms the speech into acoustic feature of log-scaled filter-banks.

[1]The video of the demo can be found online at https://youtu.be/7brmUKbyJJc

The acoustic features are also post-processed by Cepstral Mean and Variance Normalization (CMVN).

The local acoustic pattern extractor is based on a 2D Convolutional Neural Network (CNN) modified from the well-known ResNet-34 (He et al. 2016) architecture. It has 4 residual blocks, where each convolutional layer is followed by a batch normalization (BN) layer and a ReLU activation function. Taking the filter-banks as input, the modified ResNet-34 effectively extracts the local patterns with respect to the speaker identity and outputs a sequence of frame-level representations.

In order to aggregate the speaker-specific patterns in an adapted manner, the embedding layer leverages a multi-head attention mechanism (MHA) (Vaswani et al. 2017) to attend to different discriminative patterns in the speech. By concatenating the outputs from different heads of Scaled Dot-Product Attention (Vaswani et al. 2017), the MHA integrates the frame-level representations into a utterance-level representation. With another two consecutive fully-connected layers, the utterance-level representation is further transformed into the final speaker embedding. Given the speaker embeddings of input speeches, we can measure their similarities in terms of speaker identity through the cosine similarities between the speaker embeddings.

For discriminative speaker embeddings, the speaker embedding is optimized to classify speaker identity with the additive margin softmax (AM-Softmax) loss (Wang et al. 2018a,b). By forcing the cosine similarity toward the ground-truth speaker identity to be at least a predefined margin more than those toward the false speakers, the system is able to capture more discriminative and robust speaker embeddings.

The system is trained with a large dataset containing more than 40,000 speakes and the dataset is composed of a wide range of speech datasets, including VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), VoxCeleb2 (Chung, Nagrani, and Zisserman 2018), Aishell-1 (Bu et al. 2017), Aishell-2 (Du et al. 2018), MAGIDATA[2], etc. Evaluation on short duration test speeches (i.e 10 second speech for register and 3 second speech for verification) reports an Equal Error Rate (EER) of 0.076% in complex testing scenarios, showing the superior performance of the system in speaker verification.

Demonstration

The demonstration shows the application of the speaker verification system on mobile devices. The snapshots are shown in Figure 2. The main functions of the application includes two phrases: *register* and *testing*. In the register phase, the users need to register in the system with three different speeches. Specifically, they can read the given numbers and mottoes in the demo or speak any other texts they like. Each registered speech should be at least 3 seconds to ensure verification performance. The system converts the registered speeches into speaker embedding through the techniques discussed in the previous section and stores it in the server.

In the testing phase, the users can speak the given texts or any other text they like to verify their identity. And the verification decision is made by comparing the speaker embedding of the input speech with the registered speaker embedding. Particularly, the demonstration examines the scenario with mask wearing and finds that the system works as well as general scenario, showing the accuracy and robustness of the system.

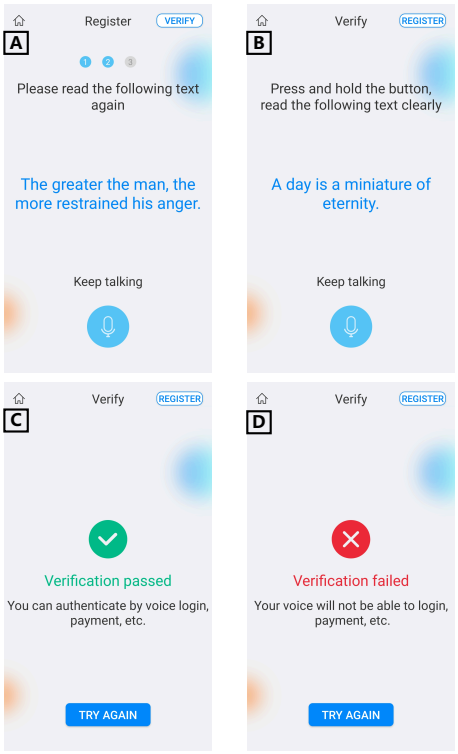

Figure 2: Snapshots of the speaker verification system: (A) Register; (B) Testing; (C) Verification passed and (D) Verification failed.

The system is deployed on a machine with 126GB memory, 32 Intel Core Processor (Xeon) and CentOS. During demonstration, the audience can experience the system through cellphone provided by us or through mini-program service of their own cellphone if they have installed the WeChat app.

Conclusion

In this demonstration, we present a health-friendly speaker verification system on mobile device. We demonstrate the superiority and robustness of the system in scenarios such as wearing mask in registration or verification phase. The demonstrated system has significant social impact under the background of coronavirus pneumonia epidemic. We wish it pave the way for better bio-metric techniques which prioritize personal health and public hygiene.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2018AAA0101100.

[2]http://www.openslr.org/68

References

Bu, H.; Du, J.; Na, X.; Wu, B.; and Zheng, H. 2017. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *O-COCOSDA*, 1–5.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*, 1086–1090.

Du, J.; Na, X.; Liu, X.; and Bu, H. 2018. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale. *arXiv preprint arXiv:1808.10583* .

Hansen, J. H. L.; and Hasan, T. 2015. Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine* 32(6): 74–99.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Li, N.; Tuo, D.; Su, D.; Li, Z.; and Yu, D. 2018. Deep Discriminative Embeddings for Duration Robust Speaker Verification. In *Interspeech*, 2262–2266.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech*, 2616–2620.

Novoselov, S.; Shchemelinin, V.; Shulipa, A.; Kozlov, A.; and Kremnev, I. 2018. Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition. In *Interspeech*, 2242–2246.

Okabe, K.; Koshinaka, T.; and Shinoda, K. 2018. Attentive Statistics Pooling for Deep Speaker Embedding. In *Interspeech*, 2252–2256.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *ICASSP*, 5329–5333.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018a. Additive Margin Softmax for Face Verification. *IEEE Signal Process. Lett.* 25(7): 926–930.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018b. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *CVPR*, 5265–5274.