EC 601
Project 1
Product Review
Yifan Wang
Instructor: Osama Alshaykh

Multi-Speaker Identification

Abstract

In 1999, the National Institute of Standards and Technology (NIST) added a new task to the evaluations which is the speaker recognition evaluation that was focusing on one-speaker detection. Since then, speaker identification has been evolving quite a lot. Nowadays, one speaker identify is not a challenging task and does not meet our needs for daily use anymore. Voice control has been widely used in several areas including smart devices voice user interface, voice access for authentication, even voice control for web browser is available. So how did these functions distinguish between different speakers, how to identify whose voice is who's during a multi-person conference or meetings, how to track people's voices and separate then when they speak during the same time period will be the main focus for this review paper to analyze and discuss.

Introduction

Speaker identification is considered as the particular speaker recognition in a segment of speech, and multi-speaker identification has three tasks that been defined as multi-speaker detection, speaker tracking and speaker segmentation. Multi-speaker detection is to determine if certain speakers is present during a speech segment, speaker tracking is to determine the recognized speakers' speaking intervals in a particular speech segment, and segmentation is to determine the unrecognized presented speakers' speaking intervals in a particular speech segment. Currently some of the conference applications are doing this multi-speaker identification in a "lazy" way, which is making the audio transcript based on the speakers' id to tribute their speech, but it is not the multi-speaker identification we are discussing here. Many research and developments relating to this topic has been published and some of the methods and approaches has already being used in many products. We will look into some existed method to recognize speaker based on analyzing their voice and matching dataset to correctly tracking and identifying multi-speakers.

Use cases and applications

One of the approaches published in 2018 by Manthan Thakker, Prachi Ved, Shivangi Vyas and Shanthi Therese S. is to first use their built artificial intelligence system to identify an audio file as a single or multi-speaker file, and then identify the speakers based on their voice characteristics. Their approach for multi-speaker identification's first step is to conduct pre-processing of the audio input file, then process the raw audio file through several procedures including reduction and silence removal, framing, windowing and discrete cosine transform (DCT) calculation. After processing through, the features should be able to extracted from the post-processing audio file by using Mel frequency cepstral coefficients (MFCC) technique to perform the extraction. In order to train the system for identification, they then feed in the extracted features, and then via the neural networks using error back propagation training algorithm (EBPTA) to train the system. In short, by using Mel frequency cepstral coefficients technique for feature extraction and error back propagation training algorithm (EBPTA) for feature matching, they achieved the desired result which is to identify speakers in a multi-speaker environment. One of the many applications of this model is in biometric systems such as telephone banking, authentication and surveillance.

The most impressive feature of this system is that since the algorithm used is neural network, which operate by mimic the network and processing of a human brain, the identification is very accurate. Also, the learning potential of this system is high and it is always capable of training and getting familiar with new datasets.

However, the disadvantage for this neural network system is also obvious, which is the lack of practical using in real-time identification. In order to achieve real-time recognition of speakers, it demands massive datasets and cluster of computers to train the system to do parallel computation. Otherwise, it can only process the frame by frame classification to identify the speakers, which means in order to achieve the processing speed, it also requires huge amount of computation, thus, it requires a powerful CPU to utilize the functions.

Similar to the approach mentioned above, there are also many two-step approaches that either trained or pre-trained the source separation and automatic speech recognition (ASR) networks separately like mentioned above, then making use of mixtures and their corresponding isolated clean source references.

Several previous works have considered a two-step procedure in which first separating the mixed speech signal, and then recognize and analyze each separated speech signal for identification. Although using trained deep neural network to match each time-frequency unit to a high-dimensional embedding vector such that the embeddings for the time-frequency unit pairs dominated by the same speaker are close to each other, while those for pairs dominated by different speakers are further away is considered as a huge improvement, and it could technically be trained without references for the isolated processed speech signals, it is still difficult to train from scratch in that case.

Another approach also published in 2018 by Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, and John R. Hershey proposed an end-to-end multi-speaker speech recognizer based on permutation-free training and a new objective function promoting the separation of hidden vectors in order to generate multiple hypotheses. In an encoder-decoder network framework, teacher forcing at the decoder network under multiple references increases computational cost if implemented naively, and they avoided this problem by employing a joint connectionist temporal classification/attention-based encoder-decoder network.

The highlight is that the performance is better compare to the previous non-end-to-end system including deep-clustering-based speech separation and a Kaldi-based automatic speech recognition system for multi-speaker speech recognition. However, it still requires the two-step procedure and the mixed-speech files they used for training the system are simulations not the real-life cases, so the practicability is questionable and there are no actual applications mentioned so far.

Discussion

Based on my research, seems like most of the existing multi-speaker identification are using the trained deep neural network to perform the voice matching and recognition, some of the existing applications and APIs such as “Deepaffects”, “Google Speech”, “IBM Watson” and so on are all using large datasets of voiceprints or speeches to pretrain the neural network via either separated speech signals or mixed postprocessing speech signals. Therefore, I would make the conclusion for now that the best way to approach the multi-speaker speech environment identification is to build the system of deep neural network and train it with numerous speech files in order to achieve the desired performance. Although it is possible to find an alternative approach, I think that the accuracy and the executability would be lower than using the deep neural network and the developing capacity would also be more limited.

Conclusion

Multi-speaker identification is a popular developing area, and with so many great researchers and developers contribute together to this topic, the analyzing technique and algorithm is more and more advanced, this will lead deeper down to the neural network which is an essential part of artificial intelligence. I am really looking forward to try to approach the goal during this exciting semester since this is an uncharted field to me and I can foresee the potential things I can learn through this pass. Since the core methods are all related to neural network based on my research, it is hard for me to try to duplicate any of the approach in such a short time period, but I believe that with awesome teamwork later on, and chasing further down into this area, I can definitely get some exciting result.

References

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In Advances in Neural Information Processing Systems (NIPS), pages 577–585.

Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(10):1901–1913.

Seki, Hiroshi et al. “A Purely End-to-end System for Multi-speaker Speech Recognition.” ACL (2018).

Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, and John R. Hershey. 2018. End-to-end multi-speaker speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4819–4823.

Thakker, Manthan & Vyas, Shivangi & Ved, Prachi & Therese, s. (2018). Speaker Identification in a Multi-speaker Environment. 10.1007/978-981-10-3920-1_24.

Yanmin Qian, Xuankai Chang, and Dong Yu. 2017. Single-channel multi-talker speech recognition with permutation invariant training. arXiv preprint arXiv:1707.06527.