Text-Independent Speaker Identification with Breath and Ey

Wenbo Zhao
School of Electrical and
Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
Email: wzhao1@andrew.cmu.edu

***Abstract*—Identifying speakers in text-independent scenarios from short period speech recordings is difficult due to rich variations of human voices, spoofing voices, noisy recording channels, etc. In prior work, to represent the variations and similarities in speech, statistical characteristics such as the Gaussian Mixture Models are used. While effective, they rely on prior assumptions of the data distribution. Moreover, they require iterative steps and large amount of data, resulting in inefficiency and poor performance when data is insufficient. To overcome these problems, in this study, it uses constant-Q spectrograms to represent the recordings. This representation is data distribution assumptions free, efficient, and works well with small amount of data. To identify speakers using only their breath or their 'Ey' sound recordings, this study further develops a pipeline using convolutional network and Long Short-Term Memory network. Compared with the state-of-art i-Vector pipeline, the proposed pipeline is simple and effective.**

I. Introduction

Speaker recognition refers to two different tasks involving discriminating people from their voices. The two tasks are *speaker verification* and *speaker identification*. Speaker verification determines whether or not a voice belongs to a person it claims to be, while speaker identification determines the speaker from a set of known speakers for a given voice. Speaker recognition can be *text-dependent* or *text-independent*, where the former means the spoken words are predefined, like a phrase, and, on the other hand, the latter means no prior knowledge of the spoken words. In this paper, we focus on text-independent speaker identification. We start by reviewing the literature.

A. Review of Speaker Identification Methods

Generally speaking, text-independent speaker identification extracts spectral characteristics or statistical characteristics from utterances, and matches them with the known speakers' characteristics to determine the speakers. We review the commonly used characteristics, and discuss the traditional approaches used to match speakers.

The most popular spectral characteristics for speech data is the Mel-frequency cepstral coefficients (MFCC). It is a good local representation of a short frame of speech spectrum. ...(short time segmentation for stationary, Mel/logarithmic frequency scale and triangular filter banks for mimicking human auditory, log for spectral separation, DCT for nulling high frequency variations, delta and second delta for velocity and acceleration, mean normalization ...)

Compared with the deterministic MFCC feature, another feature, the Gaussian Mixture Models (GMM) statistically model the power spectrum density of speech with mixture of Gaussian distributions. The reasons using mixture of Gaussians are two-folded: (1) ...(2) ...The GMM model is governed by three factors: the mixture responsibility, the mean, and the variance. In order to infer the model factors, we use Expectation Maximization (EM) method to iteratively maximize the conditional expected log likelihood of the speech data.

Another statistical characteristics for speech, the i-Vectors, also use GMM but separately model the speaker-independent feature distributions (the Universal Background Model (UBM)) as well as speaker-dependent feature distributions (the supervector). Performing subspace decomposition (joint factor analysis) on the supervector yields the i-Vectors. It ...

We have discussed the spectral and statistical characteristics for characterizing the speech data. Next, we discuss the approaches to identify speakers using these characteristics. Specifically, we analyze five different approaches: (1) dynamic time warping (DTW), (2) hidden Markov models (HMM), (3) vector quantization (VQ), (4) time delay neural networks, and (5) recurrent neural networks (RNN).

...

B. Contribution of This Paper

The above-mentioned characteristics are, generally speaking, effective in capturing the rich variations of human voices and identifying speakers. However, these statistical characteristics require large amount of data to train, has iterative steps, and rely on prior assumptions of the data distributions, thus are inefficient and perform bad when lack of data. On the other hand, the spectral characteristics, the MFCC feature in particular, are easy to compute, do not rely on large data, and store the original information from speech signals, thus are preferred in this study. But, the MFCC feature cannot leverage the pitch variations in speech.

In this study, our objectives are to develop a spectral characteristics to capture the pitch variations in speech, and develop a pipeline to identify speakers using breath and 'Ey'

sound recordings. First, the breath recordings are transformed to constant-Q spectrograms to leverage pitch variations among speakers. Then, the spectrograms are fed into convolutional network to capture their shift invariant features. Next, these features are fed into a Long Short-Term Memory network to capture the temporal variations. Finally, a feedforward network predicts the multi-class likelihoods for each recording. Compared with conventional methods, the proposed method is (1) simple, without complex feature computation, (2) effective on small amount of data, and (3) independent of assumptions of data distributions.

C. Organization of This Paper

II. The Speaker Identification Framework with Breath and Ey

A. Problem Formulation

Objectives
Explain prior ideas

B. Feature Extraction

C. Speaker Identification Network

1) Convolutional Network:

2) LSTM Network:

D. Experiments and Results

What you have done
Key technique decisions
Results

III. Conclusions

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.