Neural Networks Based 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

I. Motivation

Speaker recognition is useful in many applications, such as in voice authentication, fraud caller detection, multi-speaker tracking, etc. It usually refers to two different tasks involving discriminating people from their voices, *speaker verification* and *speaker identification*[1]. Speaker verification answers question like “Did this person speak this utterance or not?” by comparing the claimed speaker with a potentially large set of speakers, while speaker identification answers question like “Who spoke this utterance?” for a close-set of known speakers. Speaker recognition can be *text-dependent* or *text-independent*[1], 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 study, we focus on text-independent speaker identification.

Generally speaking, the speaker identification process includes three phases: feature extraction, speaker modeling, and decision making [2], [3]. We study the first two phases.

Spectral features reflect the mechanical aspects of articulatory system [1], which are unique among individuals, and hence can be used to distinguish speakers. The Mel spectrogram is one of the commonly used features [1].

In text-independent case, Gaussian Mixture Models (GMM) is one of the most commonly used speaker models. It statistically models the speech features as a finite mixture of Gaussian distributions [2]. GMM is a reasonable model because firstly, it represents each class of acoustic events with an individual Gaussian[3], and secondly, a linear combination of Gaussian basis functions can smoothly approximate a large family of densities.

Despite of many of its advantages, our study shows that Mel spectrogram is not a good representation for very short utterance in text-independent speaker identification tasks. The reason is that a good feature should give similar representations for utterances from a same speaker, and dissimilar representations for utterances from different speakers, but Mel spectrogram fails to do so in this problem settings.

On the other hand, GMM relies on prior data distribution assumptions, which may fail when these assumptions do not hold. Training these models usually requires large amount of data, but in many cases we only have small amount of short phone call recordings. Moreover, GMM has iterative procedures, resulting in inefficiency.

II. Methods

In this study, we propose an alternative approach to address the drawbacks associated with Mel spectrogram feature and GMM model.

We consider the task of identifying speakers with breath and ‘Ey’ sound (as in ‘Mayday’ as an international radio distress signal) recordings. These recordings are intercepted from phone calls. We propose to use constant-Q features instead of Mel spectrograms. We observe that constant-Q features are good representations for short utterances in our problem settings because firstly, they promote distinction between speakers, and secondly, they help leverage pitch variations within the same speaker.

Furthermore, we propose a neural networks based speaker model. It models constant-Q spectrograms using convolutional network (CNN) and Long Short-Term Memory (LSTM) network. CNN automatically learns speaker’s shift-invariant features from constant-Q spectrograms. LSTM captures the temporal information from length-varying input sequences. The CNN-LSTM model directly outputs speaker likelihoods for decision making.

III. Progress

We test our proposed method by identifying speakers using breath and ‘Ey’. There are 44 speakers in the breath task, and 53 speakers in the ‘Ey’ task. Preliminary results show that our method achieves high identification accuracy.

IV. Comparison

In text-independent scenarios and with short utterances, compared with Mel spectrograms, the constant-Q features are more discriminative. Compared with GMM speaker model, our proposed CNN-LSTM model is accurate, distribution assumption free, and fast.

References

[1] R. S. S. Kumari, S. S. Nidhyananthan *et al.*, “Fused mel feature sets based text-independent speaker identification using gaussian mixture model,” *Procedia Engineering*, vol. 30, pp. 319–326, 2012.

[2] R. Zheng, S. Zhang, and B. Xu, “Text-independent speaker identification using gmm-ubm and frame level likelihood normalization,” in *Chinese Spoken Language Processing, 2004 International Symposium on*. IEEE, 2004, pp. 289–292.

[3] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.