Digital Signal Processing: Speaker Recognition
Midterm Report

Xinyu Zhou, Yuxin Wu, and Tiezheng Li
Tsinghua University

1 Task

Build a speaker recognition system

2 Dataset

The dataset provided by teacher comprised of 102 speaker, in which 60 are female and the rest are male, with three different speaking style: Spontaneous, Reading and Whisper. A statistic is as follows: All utterances are first feed to a VAD process prior futher processing.

	Spontaneous	Reading	Whisper
Average Duration	202s	205s	221s
Female Average Duration	205s	202s	217s
Male Average Duration	200s	203s	223s

3 Approach

Based on weeks of literature reviewing and testing, we have designed our overall approach to this task as followed:

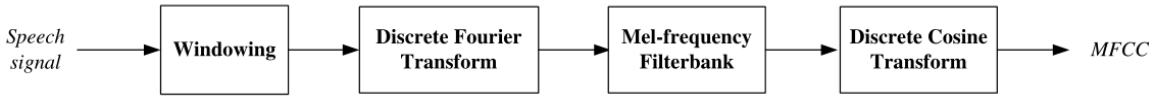
1. Energy-Based VAD

   Input audio signals are very likely to contain significantly large ratio of blank signals. VAD (Voice Activity Detector) is a preprocessing technique to filter out the blank period. The most common approach toward this goal is to use energy-based feature of signals.

2. Cepstrum-Based Features

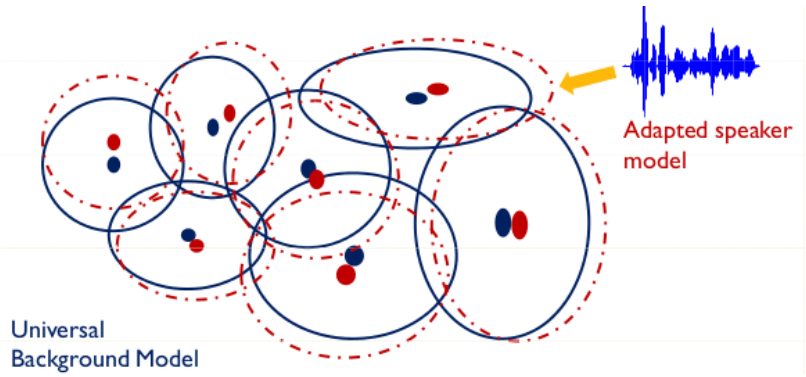
   Research showed that cepstrum-based features are more discrimitive in the task of speech and speaker recognition/verification. We decide to apply common cepstrum features extraction routine, such as MFCC (Mel-frequency Cepstrum Coefficients), LFCC, after the original signals are preprocessed by VAD.

   The basic procedure of MFCC is shown below, and the details about extracting MFCC is already been explained in the previous reports. The output of this step, is a sequence of fixed-dimension vectors, which will be used later in model training.



3. GMM and UBM Model

   Since the feature vectors of a speaker tend to cluster into **several** groups in the feature space, the model of a specific speaker can be well described by using GMM (Gaussian Mixture Model). Moreover, for different speakers, the components of their individual GMMs will also have similar distributions, which discriminate different syllables. Therefore, an UBM (Universal Background Model) can be first trained for all speakers, and then we can get adapted GMMs which fit this task better.



4. JFA GMM-based model can describe the clusterred distribution, but it fails to account for different types of variability in each clusterred group. However, we only need inter-speaker variability to be modeled, but not channel or noise variability. Join Factor Analysis can convey such information.

4 Progress

Till now, we have implemented MFCC feature extractor and GMM model for acoustic modeling. For GMM we employ scikit-learn[2].

We've tested on 20 speakers for a closed set recognition task and calculate accuracy of recognition. 30 seconds utterance is used for training, and 5 seconds utterance is used for test. We randomly extract 100 continuous 5-second test utterance from utterance of a speaker as test set. There's no overlap between training and test utterance.

For the last week, we fine-tuned parameters of our model and scrutinizing more paper preceding to our work.

The test is repeated 20 times to obtain an accurate estimate of accuracy. The test result is as follows:

- **Spontaneous** 0.940

- **Reading** 0.926
- **Whisper** 0.931

5 Analysis

The result indicates the effectiveness of our method, but also shows the limitation.

1. The performance is not satisfying.
   Compare to the performance given in serval sources (books and papers), ours are not the best. This may due to the corpus difference between test, and the result may not be comparable. But comparing to MFCC extractor provided by bob[1], we get 5% higher accuracy, which proved the effectiveness of our model.
2. Long training utterance.
   30s utterance training data may not feasible in practical application.
3. GMM is of low efficiency when classifying.
   Due to the modeling of each spkeaker using GMM with 32 mixtures, classification of a single speaker involves scoring over all enrolled speakers. The more speaker, the less efficiency.

6 Future Work

In the following two weeks, we plan to:

1. Compare performance with other algorithms on the same corpus.
   Comparison is of importance since it gives us vital feedback on our approach.
2. Reduce test utterance length.
   5s utterance after VAD can be reduced to make our approach more applicable.
3. Employ GMM-UBM to get more accurate GMM model.
4. understanding JFA.
5. Extend our method to open set recognition.
6. GUI functional design.

Moreover, we've proposed some other methods which need futher investigation:

1. using Bootstraped-Multi-GMM to convey more information.

References

[1] A. Anjos et al. "Bob: a free signal processing and machine learning toolbox for researchers". In: *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. ACM Press, Oct. 2012. URL: http://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf.

[2] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.