Digital Signal Processing: Speaker Recognition
Midterm Checkpoint Report

Xinyu Zhou, Yuxin Wu, and Tiezheng Li
Tsinghua University

1 Progress

During past weeks, we've built a proof-of-concept ASR system based on MFCC feature and GMM for modeling spearker with fine-tuned parameters.

A brief performance summary:

- $10s \sim 15s$ training corpus per person.
- $2s \sim 5s$ test corpus per person.
- We adopted GMM with 32 Gaussians.
- Accuracy on 5 speakers is about 93 percent
- Accuracy on 10 speakers is about 90 percent

It turns out that, our method worked well when the condition that the number of speakers is limited to 5 or less is met. But as we employed the new dataset which provided by teacher this week, which contains 102 speaker, comprised of 60 females and 42 males, in three speaking conditions: reading, spontaneous and whisper, the challenge we are facing is beyond our expectation.

Although we fulfilled the plan we aforementioned in opening report, the overall performance is still unsatisfying. The main reason that MFCC + GMM approach suffers from new corpus is that, using solely MFCC limits the model we can use to generative models, which in turns brings computation inefficiency to feature extraction, model training and testing.

2 Next Step

Our proposed way to workaround the situation mentioned in previous chapter is using supervector-based approaches. This is advantageous since it map a speech to a single vector, which makes it easy to utilizes discrimitive models that are more powerful in classification task, such as SVM. Further more, we plan to conduct extensive investigation on **Joint Factor Analysis(JFA)** , which is a much promising method to our best knowledge extent.