Speaker Recognition

Notes for Automatic Speech Recognition

Chia-Ping Chen

cpchen@cse.nsysu.edu.tw

Department of Computer Science and Engineering

National Sun Yat-Sen University

Kaohsiung, Taiwan ROC

Introduction

- Speech also contains information about the speaker, such as identity and mood, in addition to the linguistic content (words).
- Automatic speaker recognition uses the many of the same tools and ideas of ASR.
- There are two common tasks in speaker recognition
  - verification: decide if the speaker is the one as he/she claims to be (yes/no question).
  - identification: classify the speaker as one of a fixed set of candidates (choose-one question).
- Applications: access controls, particularly remote access.

Calculating Scores

- Since we are mapping a speech to a speaker, we need a “score” to reflect the distance between an utterance and a candidate.
  - One simple method is to represent a speaker by a Gaussian mixture in the space of speech features.
  - A speaker can also be represented by HMMs
    - fully connected for text-independent tasks
    - left-to-right for text-dependent tasks.

Acoustic Parameters

- In speech recognition, we want features in speech that are
  - indicative of the words while invariant to the speaker
  - robust to noises
- In speaker recognition, we want information in speech that is
  - idiosyncratic of the speaker while independent of the words
  - robust to noises
- Surprisingly, current speaker recognition systems often use the same parameters (MFCC) as the ASR.

Statistical Framework

- Using the statistical pattern recognition approach, the posterior probability for a candidate speaker S_c given the acoustic parameters X is

$$p(S_c|X) = \frac{p(S_c)p(X|S_c)}{\sum_i p(S_i)p(X|S_i)}.$$

- For speaker identification, we use MAP criteria, and

$$S^* = \arg\max_S p(S|X) = \arg\max_S p(X|S)p(S).$$

Speaker Verification

- For speaker verification, we accept the hypothesis

$$S = S_c \text{ if } \frac{p(S_c|X)}{p(\bar{S}_c|X)} > \delta,$$

where $\delta > 1$ and $p(\bar{S}_c|X) = 1 - p(S_c|X)$.

- Assuming that the candidate set is exhaustive and has uniform prior probability, then

$$p(\bar{S}_c|X) = \sum_{i \neq c} p(S_i|X),$$

$$S = S_c \text{ if } \frac{p(S_c|X)}{p(\bar{S}_c|X)} = \frac{p(X, S_c)}{\sum_{i \neq c} p(X, S_i)} = \frac{p(X|S_c)}{\sum_{i \neq c} p(X|S_i)} > \delta.$$

Cohort Set

- Using the $\log$ likelihood ratio, we have

$$S = S_c \text{ if } \log\, p(X|S_c) - \log \sum_{i \neq c} p(X|S_i) > \Delta.$$

- It is expensive to compute the sum. We can define a cohort set R_c for a speaker S_c, and approximate by the sum of the cohort set.

$$\log\, p(X|\bar{S}_c) \triangleq \log \sum_{i \neq c} p(X|S_i) \sim \log \sum_{S_i \in R_c} p(X|S_i)$$

Comments on Approximations

- It has been shown that including S_c in the cohort set R_c is beneficial. Specifically, it improves in the cases where the acoustics of utterance are rather distant from S_c, resulting a small and reliable likelihoods for both cohort and candidate.
- When HMMs are used for computing scores, one can approximate $p(X|\bar{S}_c)$ by a speaker-independent model.

Approaches for Speaker Verification

- **Text-dependent**: an user is allowed only to speak certain words. Such a system knows in advance the words to be uttered.
- **Text-independent**: an user is allowed to say any words. In such a system the lexical content of verification utterance can not be predicted.
- **Text-prompted**: an user is instructed to speak the prompted words. Such a system reduces the risk of frauds since otherwise the speech can be prepared in advance such as recording or editing from the true speakers.

Optimal Threshold

- There are two kinds of errors in verification tasks
  - false rejection: rejecting an enrolled user.
  - false acceptance: accepting an imposter.
- If the threshold is too high, the number of false rejections will increase. If the threshold is too low, the number of false acceptances will increase.
- For system comparison, the equal-error-rate threshold, where the false rejection rate equals the false acceptance rate, is used. EER can be approximated by one-half the sum of the two error rates.