Selected Applications

Notes on Speech and Audio Processing

Chia-Ping Chen

cpchen@cse.nsysu.edu.tw

Department of Computer Science and Engineering

National Sun Yat-Sen University

Kaohsiung, Taiwan ROC

Speaker Recognition

- Speaker recognition is the problem of classifying of a speaker's identity from an acoustic signal.
- It uses many of the same tools for automatic speech recognition.
- It requires a comparison of a speech utterance to a set of references.
- The identity of a speaker is evidenced by things such as pitch range, dialect, gender.

Pros and Cons as a Biometric

- Pros
  - non-intrusive
  - low-cost
  - cannot be forgotten or stolen
  - can be used remotely
- Cons
  - accuracy

Speaker Recognition Problems

- 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.
- Commercial systems are already in use.

Calculating Scores

- Since we are comparing a speech with a set of speakers, we need a “score” to reflect the distance between an utterance and any candidate.
  - One simple method is to represent a speaker by a Gaussian mixture in the speech feature space.
  - 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
  - indicative of the words while invariant to the speaker
- In speaker recognition, we want information
  - idiosyncratic of the speaker while independent of the words
- Surprisingly, speaker recognition systems often use the same parameters as the ASR. LPC-based feature sets are favored over MFCC-based.
- Robust to channel distortion is an important issue.

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 the total sum by the sum of the cohort set.

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

Comments on Approximations

- Including S_c in the cohort set R_c is advantageous. It helps in the cases where the acoustics of the actual speaker is rather distant from the claimed speaker identity S_c, resulting in very small and unreliable 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 M,

$$\log\ p(X|\bar{S}_c) \sim \log p(X|M).$$

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.

Text-Independent Approach

- Long-term statistics, such as the mean and variance over a long sequence in various domains.
- Vector Quantization. A speaker is modeled by a sequence of codebook entries.
- Fully-connected HMMs. Each speaker is modeled by a fully connected HMM.
- ANN. Each speaker corresponds to a neural net.

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.

Speech Synthesis

- A “talking machine”
  - domain: what can it say?
  - method: how does it say something?
- Basic steps
  - text processing: resolving ambiguity in the text
  - processed text to phonetic-prosodic translation: determining sound units, pitch, duration and amplitude
  - speech generation

Text Processing

- Editing: “hte” → “the”
- Acronyms: “a.k.a. TTS” → “also known as text to speech”
- Abbreviations: “St.” can be “street” or “saint”
- Numbers: “10” can be “one-zero” or “ten”
- Symbols: $ (dollar), % (percent), @ (at)
- Dates and times

Phonetic and Prosodic Translation

- Phonetic part: uses dictionary lookup and pronunciation rules. First determine the canonical pronunciation of each word to be synthesized and then apply the pronunciation rules.
- Prosodic part: parses the input text for the syntactic structure, then apply rules to assign prosodic features to parts. This is the more difficult part.

Speech Generation

- Articulatory synthesis: directly modeling the physical system for articulators and their movements. Parallels the mechanical systems, e.g. by von Kempelen.
- Source-filter synthesis: simpler than the articulatory synthesis. The sound signal is modeled by filter(s) driven by source(s).
- Concatenative synthesis: stores waveforms for units, such as diphones or demisyllables, and concatenate them.

Formant Synthesizer

- OVE II by Fant et al.
  - 3 parallel branches. Top branch is used to synthesize vowels, semivowels and glides. Middle branch is for nasals. The bottom branch is for fricatives and plosives.
  - 2 excitations. Pulse generator is used in vowel and nasal. Noise generator is used in whispered vowels or fricatives.
  - Most of the sounds can be synthesized by balancing the parameters.
- See also Holmes and Klatt for variations.

Concatenative Methods

- Speech segments (units) are collected from speech corpus via methods such as forced alignment.
- Each unit is labelled by its pitch, duration, amplitude and the identity of neighboring units.
- Given the prosody-tagged text, the database is searched for the optimal sequence of units.
  - Algorithms such as dynamic programming are used for efficient search.
  - The design of cost functions is a research issue.