Human Speech Recognition

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- How do people recognize and understand speech? For this question, much has been said, but little has been understood or agreed.
- We can only hope to introduce a few key concepts. The difference between human recognition and artificial recognition will be emphasized.
- We focus on two studies.
  - The perception of CVC syllables.
  - The comparison of human and machine performance on speech recognition tasks.

Allen’s Model

- Humans do partial recognition of phonetic units across time, independently in different frequency ranges.
- This suggests a subband analysis for speech recognition. Subband informations are integrated at the level of phonetic categorization.
- Note that this is at odds with current ASR systems which are almost all based on frame-wise short-term spectral estimates.

Articulation Experiments

- Here, the word *articulation* means the probability of correctly identifying non-sense speech sounds.
- Databases are designed from CVC, CV and VC non-sense syllables. They were believed to be an ideal testbed for speech recognition without other factors such as multisyllabic structures.
- Listening tests were conducted with varying SNR and frequency ranges (via filters).

Some Results

- The probability of getting a CVC syllable correct was roughly the product of getting the initial C, the V, and the final C correct in the syllable recognition. This means phone recognitions could be treated independently.
- The phone error probability with total spectrum was equal to the product of the error probabilities with low-passed and high-passed spectra.

Articulation Index

Let $s(a, b)$ be the articulation (probability of correct phone recognition) using the band (a, b), then

$$[1 - s(a, c)] = [1 - s(a, b)][1 - s(b, c)].$$

If we define articulation index

$$\mathrm{AI}(s) = \frac{\log_{10}(1 - s)}{\log_{10}(1 - s_{\mathrm{max}})},$$

where s_{max} is the maximum articulation, measured to be 0.985, then

$$\mathrm{AI}(s(a, c)) = \mathrm{AI}(s(a, b)) + \mathrm{AI}(s(b, c)).$$

Speech Corpora

- A speech corpus is a collection of speech data.
- For statistical approach, data is king.
- A speech corpus is characterized by
  - style (read, spontaneous, isolated)
  - no. of talkers
  - vocabulary size
  - no. of utterances
  - data size (duration)
  - recognition perplexity
- See Table 18.1 (or simplified Figure 18.1) for some examples.

HSR and ASR

- Although making progress, ASR has much work ahead to catch up with HSR.
- The extent of HSR superiority increases with the difficulty of a recognition task.
  - noisy speech
  - spontaneous speech
- It appears that HSR is quite different from ASR in
  - signal processing and representation
  - subword recognition
  - temporal integration
  - integration of higher-level information