Speech Perception

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 human perceive speech is of great interest to researchers of speech processing.
  - For speech synthesis or speech coding, it provides valuable information for quality and intelligibility.
  - For speech recognition, it provides valuable information for noise robustness, since it is a working example of noise-robust system.
- “Perception” is the relation between physiological response and the mental state change of the listener.

Vowel Perception

- We have seen that different vowels have different vocal tract configurations, which lead to different formants. So there is a relation between formants and vowel perception.
- The lower formants are easier to resolve as the auditory filters have narrower bandwidth. The higher formants are harder to resolve.
- This leads to the conjecture that the perception of vowels can depend on just two effective formants.

Sachs Experiments

- Sachs et. al. study the simultaneous responses of hundreds of peripheral auditory nerves to the same stimulus. The idea is represented in Figure 17.1.
- They study the responses to “e” (as in *bet*) by computing post-stimulus time histogram (PSTH) and Fourier transform.
- For the neuron with CF $= 400$ Hz, results in Figure 17.2 show synchrony to the fourth harmonic.
- Different neurons synchronize with different harmonics, as shown in Figure 17.3.

Consonant Perception

- Early work on speech perception focused on the vowels. This is natural as properties of vowels are relatively well understood.
- However, consonant perception is even more important since consonants typically play a greater role in human understanding of speech.
- While formants are important to the perception of vowels, the formant transitions are important to the perception of consonants.

Confusion Matrix

- In Miller and Nicely, transitions into the vowel “a” (as in *father*) are experimented. The listeners are asked to decide which one in 16 different consonants is said.
- Results are compiled into a (confusion) matrix.
  - Each row represents a consonant that is heard
  - Each column represents a consonant that is said.
  - The (i, j) element of the confusion matrix is the number of times the jth consonant is recognized as the ith consonant.
- The resultant matrix has a block structure.

Consonant Recognition

- Different noise conditions are experimented.
  - Figure 17.4: SNR $= 12$ dB.
  - Figure 17.5: SNR $= -6$ dB.
  - 17 conditions are listed in Figure 17.6.
- As noise level increases, the confusion matrix becomes more “scattered”.
- Block structure in the confusion matrix of consonant recognition is related to sound features, which we describe next.

Sound Features

- Human sounds can be categorized by *features*. Miller and Nicely used the features of voicing, nasality, affrication, duration and place of articulation.
  - Voicing, nasality, and duration are self-evident.
  - Affrication is characterized with open vocal cord with a constriction.
  - Place of articulation is the position of a critical spot in the vocal tract.
- Can you identify the features for the block structure?

Binary Distinctive Feature Set

- This set is created to encompass all languages.
- There are two category of features: place of articulation and manner of articulation.
- The feature values are binary numbers.
- Each consonant is represented by 12 binary values. This is shown in Figure 17.7.

Articulatory Categories

- Consonants can also be classified by the articulatory categories, as shown in Figure 17.8.
- A consonant is defined by the required vocal tract shape, including the role of glottis, and the place of greatest constriction in the vocal tract.

Cues for Unvoiced Plosives

- For plosive sounds there are 3-4 acoustically distinct intervals: *closure*, *burst*, [*aspiration*, if voiceless] and *formant transition* into following vowel.
- The voice onset time (VOT) is the time period between the release of closure and the start of following vowel. VOT is related to voiced/unvoiced stop (plosive) perception.
- It has been shown that there are other cues for plosive perception such as the burst frequency and the identity of the following vowel.

Studies for Voiced Plosives

- Study on the response of cat's ears to /da/.
  - Use PSTHs of a large collection of fibers.
  - A synchronization measure called ALSR is used.
- Figure 17.10 shows the stimulus. Figure 17.11 shows the smoothed spectra and the ALSRs.
- One can see that temporal patterns (ALSR) is a good representation for stimulus spectrum.

Motor Theory of Speech Perception

- When incoming speech is transformed into a neural pattern, how does the brain interpret this?
- Motor theory: Our brains interpret the neural patterns based on the neural patterns we produce to articulate the same incoming speech.
- Analysis by synthesis.