Auditory Perception

Notes on Speech and Audio Processing

Chia-Ping Chen

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

Human Auditory System

- Human auditory system is capable of
  - distinguish subtle difference in frequency and intensity of sounds.
  - making sense of noisy sounds.
  - discerning mixed sounds from different sources.
  - “sounds who’s speaking!”
- Helmholtz proposed that the auditory nerve processes sound *tonotopically*, meaning different nerve bundles are sensitive to different frequencies.
- In this view, the auditory system is seen as a (sophisticated) filter bank.

The Peripheral Auditory System

- There are three components: the outer, middle and inner ears.
- The input to this system is an acoustic signal and the output is a collection of neural spikes into the brain.
- The physiological process

  air pressure change $\longrightarrow$ vibration of ear drum

  $\longrightarrow$ malleus $\longrightarrow$ incus $\longrightarrow$ stapes (middle ear)

  $\longrightarrow$ oval window $\longrightarrow$ fluid $\longrightarrow$ basilar membrane (cochlea)

  $\longrightarrow$ hair cells $\longrightarrow$ neural spikes

Cochlea

- Refer to Figure 14.3. The shape of cochlea resembles a snail. It is “unwinded” in Figure 14.4. The end points are called the **base** and the **apex**.
- The basilar membrane is relatively narrow and stiff near the base; it is wider and less stiff near the apex. As a result, high frequencies excite basal portion and die out further down, while low frequencies reach peaks near the apex.
- This mechanism supports the idea of filter bank. This is illustrated in Figure 14.5.

Auditory Nerves

- The motion of basilar membrane causes stereocilia motion, which in turn leads to neural spiking of auditory nerves.
- Refer to Figures 14.6 - 14.8 for further information on auditory nerves.
- Physiological measurements have uncovered some general properties of auditory nerves, including adaptation, tuning, synchrony, and non-linearity.

Adaptation

- When a stimulus is firstly applied, the spike firing rate of an auditory nerve rapidly increases. If the stimulus remains, the rate decreases in an exponential manner to a steady-state value.
- After the stimulus is removed, the firing rate decreases rapidly below the spontaneous rate before resuming this rate.
- Figure 14.9 illustrates this point.

Tuning

- Basilar membrane (BM) acts like a bank of tuned filters. Furthermore, the tuning frequency is a function of the position on BM.
- As a result, the auditory nerve has a similar tuning property. That is, different nerves are tuned to different frequencies.
- Figure 14.10 shows the sound pressure level that is required to increase the firing rate by a constant for 6 different nerve fibers. Here each curve represents an auditory nerve and one can see the dependence on frequency of the firing rate.

Synchrony

- Apply a tone and measure the histogram of time intervals between adjacent spikes of auditory nerve.
- The time between peaks (of histogram) is the inverse of the frequency, indicating that the spikes tend to occur in synchrony (a.k.a. phase locking) with the applied stimulus.
- It has been found that phase locking does not occur above 5 kHz, for cats.

Non-linearity

- Saturation: the rate of spikes that a nerve fiber can generate is limited biologically. Note this saturation rate depends on the fiber.
- Two-tone suppression: the application of another tone reduces the steady state firing rate of nerve caused by an original tone.
- Masking by noise: The response of a fiber to a tone is suppressed by an accompanying noise.
- Combination tones: If a fiber is excited by two tones, a combination tone not present in the stimulus may appear.

Summary

- The outer ear terminates at the eardrum. It is basically an acoustic tube.
- The middle ear transmits mechanical energy from malleus (driven by the eardrum) to stapes (driving the inner ear fluid).
- The inner ear contains fluid, basilar membrane and hair cells.
- The BM motion is transmitted to the hair cells, causing auditory nerves to fire neurons.
- Auditory nerves manifest adaptation, tuning, synchrony, and non-linearities such as saturation, masking and suppression.

Psychoacoustics

- Psychoacoustics is the science in which we quantify the human perception of sounds. We aim to derive a quantitative model that explains the results of auditory experiments.
- We relate psychoacoustic phenomena to physiological or physical measurements.
- Roughly speaking, a perceptive variable has its corresponding objective variable. But it’s a little more complicated than that.

Objective and Perceptual Variables

- Objective variables can be measured by instruments.
- For example, frequency can be measured by zero-crossing rate. Spectrum can be defined by Fourier transforms.
- The value of a perceptual variable can only be determined via psychoacoustic experiments. It is subjective.
- An objective variables has a corresponding perceptual variable.

Correspondence

- An approximate correspondence between objective and perceptive variables is

$$\begin{cases} \text{frequency} \Leftrightarrow \text{pitch} \\ \text{intensity} \Leftrightarrow \text{loudness} \\ \text{spectrum} \Leftrightarrow \text{timbre} \end{cases}$$

- The relation is non-linear: if the intensity is doubled, a subject does not “percept” a doubled loudness.
- In reality, the measure of a perceptive variable can depend on more than a single objective variable.

Psychoacoustic Tests

- It is natural to ask the following questions:
  - How does ear respond to different **intensities**?
  - How does ear respond to different **frequencies**?
  - How well does the ear focus on a given sound of interest in the presence of interfering sounds?
- Such questions can be quantitatively addressed by conducting psychoacoustic tests.
- The design of psychoacoustic tests, as well as the explanation of the results, is a tricky matter.

Sound Pressure Level

- Sensation level often varies logarithmically with the stimulus.
- The sound pressure level of pressure p is defined by

$$\text{SPL} = 20 \log \frac{p}{p_0} \text{ dB},$$

where $p_0 = 2 \times 10^{-5} \frac{\text{N}}{\text{m}^2}$ is the threshold of hearing at 1000 Hz.

Loudness and SPL

- An empirical relation is found on loudness:

$$S \propto p^{0.6} (\propto I^{0.3}).$$

  - It is a **cubic-root** law.
  - The proportional constant depends on frequency.
  - Human ears are most sensitive at 4 kHz.
- The equal-loudness curves, as shown in Figure 15.1, display the dependence of intensity on frequency for given loudness levels.
- A loudness measure is *phon*, defined to be the SPL in decibels at 1000 Hz.

Loudness and Duration

- Experiments show that if the duration of a sound is less than 200 ms, then it will appear less loud than if it were longer than 200 ms.
- Figure 15.2 illustrates this point. Basically, the minimum audible level of intensity increases as the duration decreases below 200 ms.
- Apparently, some form of temporal integration is at work when perceiving loudness.

Critical Bands

- Critical-band experiment
  - have a tone audible in band-limited white noise.
  - decrease the tone intensity until it is inaudible. Record the tone intensity.
  - repeat with a reduced bandwidth of noise.
- It is found that when the band-width is reduced beyond a critical value (called critical band), the listener's response monotonically increased. So what counts is the SNR within the critical band.
- The critical band is larger for higher-frequency tone, implying that the human auditory filters have greater bandwidths with higher center frequencies.

Masking

- When two tones are presented simultaneously, the weaker one may be masked (not heard).
- Closer tones have a greater masking effect, and a louder tone affects tones further away.
- A tone masks a higher tone easier than a lower tone.
  - Figures 15.6 and 15.7 illustrate the **asymmetry** of masking. Here the target and masker signals are reversed and the number of times by which the lower tone masks the higher tone than the converse is shown.

Summary

- Loudness is roughly proportional to the cubic root of sound intensity.
- Loudness is dependent on frequency, as demonstrated by the equal-loudness curves.
- Loudness is dependent on the duration below 200 ms, implying some kinds of temporal integration.
- Critical-band filtering of the auditory systems does exist.
- Tones mask one another in an asymmetrical way.

Pitch Perception

- Human pitch perception is performed by the entire auditory system, of which our knowledge is still fragmentary.
- The modeling of pitch perception is primarily based on psychoacoustic experiments, at times also supported by physiological explorations.
- Explanations can be controversial.

Theory of Pure Tone Perception

- How is a pure tone perceived?
- The basilar membrane responds to different frequencies with peaks at different locations. So different hair cells respond to different frequencies.
- A pure tone would cause the greatest vibration at a specific place on the BM. This ultimately leads to the perception of this tone: the brain knows which fiber(s) is excited and is able to perceive the tone.

Pitch Perception

- While the previous theory (due to Helmholtz) explains the perception of pure tones, it cannot explain the following pitch perception.
- A pitch can be perceived when the energy at the corresponding frequency is completely absent.
  - Experiment by Seebeck: The pitch at frequency $\frac{1}{T}$ can be perceived even the spectrum has no energy there.
  - A modern version: The perceived pitch remains as 200 Hz even when the pure tones at that frequency is removed.

Periodicity Model

- A periodicity model is shown in Figure 16.9. The BM is modeled as a filter bank and the hair-cell auditory-nerve complex is modeled as elementary pitch detectors (EPD).
- Each EPD fires neuron independently. A global unit accumulates the intervals between spikes. The pitch period is determined by choosing the peak of histogram.

Place Model

- The underlying assumption of the place model is the ability of the auditory system to resolve harmonic peaks of the stimulus.
- There are two stages in this model:
  - Stage 1 performs statistical separation of the frequency spacing between spectral peaks.
  - Stage 2 computes the correlation of the spectrum of lags proposed by stage 1. The winning candidate with maximum correlation is the pitch.
  - This is shown in Figure 16.10.

Speech Perception

- 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.

Human Speech Recognition

- 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

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

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

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

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