SLHS 1301 Study Guide

Chapter 9: Digital Processing of Speech Signals

1. What is "The Digital Age"? What does digital mean?

"The Digital Age" refers to the time period in the history of civilization (starting in the 1950s) in which information is recorded and processed using digital devices. Digital here means representation of information by using numbers or numerals.

2. Give some examples of digital devices. What makes them special? For example, how does a digital camera differ from a film camera?

Examples: digital phone, digital clock, digital radio, CD, DVD, digital camcorder, ipod, PDA, scanner, GPS system, voice recognition system, MP3 players, laptop computers,…

They are special in the sense that they all use numbers to represent the information (audio, video, etc.). Digital cameras decompose and encode the colors and shapes in pixels using numbers. Film cameras use lighting effects directly printed on the analog negative film which has sensitive chemicals in response to light.

3. Why is digital representation of information important?

Because digital representation completely changed the way we live our lives. Its importance lies in the precision, reliability, speed, low cost, and small size of electronic systems that perform digital manipulations.

4. What is a digital signal processor (DSP)? What kind of mathematical computation does it do to process speech and audio signal?

DSPs are special purpose Very Large Scale Integrated Circuit (VLSC) devices designed to process audiovisual and electromagnetic signals. The mathematical methods include amplification, filtering, spectrum analysis, automatic synthesis and recognition. The DSPs typically perform hundreds of millions of operations per second.

5. What is sampling? What is sampling rate? What does the Sampling Theorem say about sampling?

Sampling is a mathematical selection of data points in the signal periodically in time. Sampling rate is also known as sampling frequency, i.e., the number of samples per second.
The Sampling Theorem states that the digitized samples must be taken such that Fs ≥ 2 Fn (Fs = sapling frequency; and Fn = Nyquist frequency, which refers to the highest frequency component in the signal. If the Fs < 2 Fn, the sampling data points are not sufficent to reconstruct the frequency components in the signal faithfully. In this situation, addition of erroneous components to the signal during reconstruction may happen. This is known as

aliasing.

6. What is meant by "real time" processing?

"Real time" processing refers to nearly no time lag in processing and representing the information of a complicated system (such as tracking neural processing of speech in the brain or automatic translation of speech from one language into another as the sentences are spoken). "Real time" technology requires special computer architecture such as parallel processing to do fast-speed operations.

7. What does a digital filter do the signal?

A digital filter limits the frequency range of the signal for analysis or further processing.

8. What is spectrogram? What can it be used for? What mathematical technique does it rely on?

A spectrogram is a plot of frequency components in the signal as a function of time. In the spectrogram, the relative amplitudes of the frequency components are expressed in varying degrees of shaded darkness – the darker, the more energy it represents. It is used for analyzing the speech and nonspeech sounds in terms of their spectral and temporal features. The mathematical techniques involve Fast Fourier Transform (FFT) for spectral analysis over a very small time interval of the signal. Linear Predictive Coding (LPC) is also used for efficient computation of the spectral analysis based on the continuity and predictability of variations in the signal.

9. Suppose that you have been selected to be one of the three finalists on a real-life TV show for a billion-dollar-prize competition. You are required to come up with some innovative ideas for the greatest invention of the next millennium that uses digital spectrum analysis. What would be your proposal?

??? a billion-dollar secret. ☺

True or False:

10. Digital signals are always better than analog signals.

F.

11. Quantization error refers to the difference between the digital signal and the sample values in digital processing.

T.

12. One goal of speech coding and speech compression is to achieve high quality sound and music at higher sampling rates and with more bits more sample in order to improve the quality and fidelity of the recordings.

F.

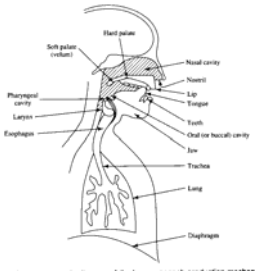
13. Human speech has almost no frequency components above 7 kHz. Therefore, digitization of speech signals needs a sampling rate of at least 14 kHz for high quality recording.

T.

14. Telephones typically use a bandwidth of only 3.2 kHz. Since human speech has a bandwidth of approximately 7 kHz, at least 50% of telephone speech is unintelligible.

F.

SLHS 1301 Study Guide



Chapter 4: Speech Production

1. What are the vocal organs for speech production?
   (Suggestion: Don't just memorize, but try to visualize the whole system by drawing a rough diagram and labeling each part.)

The principal vocal organs include: the lungs, the trachea, the larynx, the pharynx, the nose, the jaw, and the mouth (consisting of the soft palate, the hard palate, the teeth, the tongue, and the lips). These organs together form the speech production system.

(Try to do the following exercises: 1. Close your eyes, do inhalation and exhalation in slow motion. 2. Try articulating /a/, /ba/, /ta/, /ga/, /shi/, /ma/, each in slow motion, to feel how the speech production system works. 3. After that, open your eyes and try repeating the articulations in front of a mirror to watch and feel how the articulators get involved.)

2. Is the so-called speech apparatus solely devoted to speech production? (Think about the functions for various components of the system.)

No. Other main functions include breathing, chewing, facial expressions (moving the jaw, for example), etc. We can also use the speech apparatus to make sounds that are not really speech at all (for example, mimicking various kinds of environmental sounds and animal calls.)

(Note: Try to think outside the box, and envision us in the ecological system and in the whole universe. You will get a better appreciation of the amazing capacity of our speech apparatus. For instance, why do we make speech sounds at all? How did we acquire the ability to speak a language? Do other species (for example, chimpanzees) have a similar speech production system?)

3. What is the vocal tract? What are articulators? What doe the articulators do to the air stream in the vocal tract?

The part of speech production system that lies above the larynx is called vocal tract. It consists of the pharyngeal cavity (or pharynx), the oral cavity (or mouth), and the nasal cavity (or nose).

The speech articulators refer to the movable parts of the speech production system that we use to manipulate and produce various speech sounds. They are: the soft palate, the tongue, the lips, the teeth and the jaw.

The articulators are involved in creating various configurations (size and shape) of the vocal tract, which result in different resonant frequencies of the vocal tract. When the air stream passes through these special configurations, different classes of vowel sounds (in terms of tongue height and backness) and consonant sounds (in terms of various amounts of constrictions, which are also called manner, position, and voicing) are made.

4. How do stops, fricatives, approximants and nasals differ from each other in terms of articulation?

These consonants differ from each other in the manner of articulation. Stops (also called plosives) have maximal constriction in the vocal tract (complete blockage of the air stream and sudden release). Fricatives have very narrow opening between the tongue and the palate, causing turbulent noise rushing through the vocal tract. Approximants have wider openings in the vocal tract and have resonant qualities close to vowels. Nasals are unique in that the soft palate is lowered and the air stream can pass through the nose. For non-nasal sounds, the soft palate is raised closing the nasal cavity so that the air stream can only pass through the oral cavity.

5. What are the main articulatory differences between consonants and vowels?

The main articulatory differences between consonants and vowels are as follows:
1. The amount of constriction between articulators (vowels allow easy sustainable passage of the air stream in the vocal tract whereas consonants do not).
2. The use of the tongue body (vowels are adjusted by changing the relative height and backness of the tongue. For some vowels, lip rounding also plays an important role. By contrast, consonants rely on the coordination of articulators to create varied degrees (manner) of constrictions in different places. In addition, we can produce either voiced or voiceless consonants by manipulating the vibrations of the vocal folds.

6. What are the main acoustic differences between consonants and vowels?

The main acoustic differences between consonants and vowels lie in the spectral and temporal characteristics. Vowels have long durations and are salient in acoustic energy with distinct formant frequency (resonances of the vocal tract where the acoustic energy is concentrated) patterns. Consonants are much shorter in duration in normal speech, and do not have steady formant frequency structures. In a syllable consisting of C (consonant) and V (vowel), the C would show some formant transition, reflecting the articulatory adjustment in the vocal tract to create the resonances for the following vowel.

7. Explain why whispered speech can be less intelligible compared with normal speech.

First, whispered speech is lower in terms of acoustic energy compared with normal speech. Second, the prosody of the speech carried by f0 is much harder to convey in whispered speech. Third, whispered speech makes it hard to tell the distinction of voiced and voiceless sounds (such as /b, d, g/ vs. /p, t, k/).

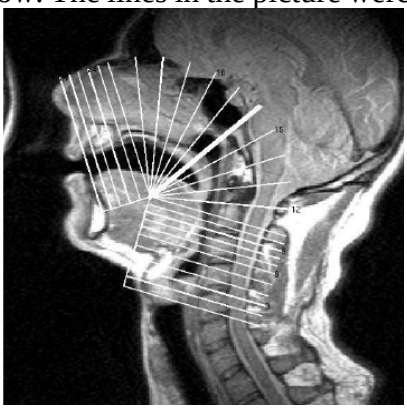
8. Explain how vocal folds are set into vibration and how the cardinal vowels in English are made (Again, try to draw the larynx, vocal folds, and some outlines of the vocal tract for the different vowel articulations.)

Please refer to pages 51-58 for a complete description. The main steps are:
1. Vocal fold closure, and manipulation of the diaphragm, abdominal muscles, rib cage muscles,

and the lungs for the right amount of pressure buildup below the larynx
2. Manipulation of the cartilages and muscles of the larynx to let the air push through the vocal folds (abduction)
3. Bernoulli effect (adduction; suction in the narrow part of the air passage, that is, the glottis).
4. Pressure buildup below the larynx again, leading to the next round of abduction (step 2).
5. The next round of adduction (step 3).
The system repeats steps 2and 3 in cycles, creating a basic vibratory pattern with a fundamental frequency called f0.

Please refer to Figure 4.13 on Page 67 to describe the articulation of cardinal vowels. For example, /i/ is a front high vowel, meaning the front part of the tongue is raised to a relatively high position in the vocal tract. /u/, on the other hand, is a back high vowel with additional lip rounding. /a/ is a low back vowel.

Note: To help you visualize the articulation, see the magnetic resonance image (MRI) for the articulation of the vowel /a/ below. The lines in the picture were drawn for computer modeling.



9. Draw the sound waveform, spectrum, and spectrogram for the vowel /i/ produced by a female speaker with f0 = 200 Hz. Label the horizontal and vertical axes.

Waveform: Please check the waveform graph in Figure 4.17 on Page 75. In labeling the T (period) = 5 ms. Vertical axis is labeled amplitude, horizontal axis time.

Spectrum: Please check the spectrum graph in Figure 4.16 for vowel /I/. The spectrum for /i/ is very similar to that for /I/. Vertical axis is labeled amplitude, horizontal axis frequency.

Spectrogram: To be covered in Chapter 7 in detail (Please have a preview of Figure 7.6 on Page 149). Vertical axis is labeled frequency, horizontal axis time (shading in darkness represents amplitude).

10. True or False: Formant frequencies are simply the amplified harmonics of the vocal buzz produced by the vocal fold vibration.
F.

11. True or False: Clicks sounds are similar to plosives in that they require maximal constriction in the vocal tract, one main difference being that clicks are made during

inhaling instead of exhaling.
T

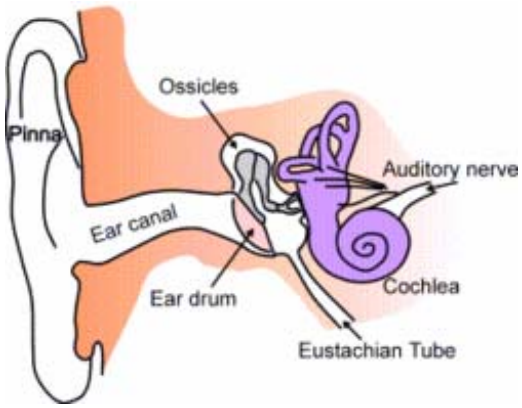
12. True or False: Trachea and esophagus are two pipes that directly connect to the stomach.
F

13. True or False: All speech sounds are produced by vibrating vocal folds.
F

14. True or False: Voiceless sounds refer to the sounds that we cannot hear.
F

15. True or False: The vocal tract serves as a resonator/filter by changing the harmonic frequencies (not the amplitude) of the buzz sound produced by the vocal folds.
F

SLHS 1301 Study Guide
Chapter 5: Hearing



1. What are the two aspects of hearing that we are interested in studying?

- the reception of sound in terms of the anatomy and physiology of the hearing organs.
- the perception of sound in terms of the sensations we experience.

2. Describe the anatomy of the auditory system.

Please see Figure 5.1, 5.2, 5.4, and 5.5 for details.

3. Describe the functions of the outer ear, middle ear and inner ear.

Out ear: sound collector and acoustic resonator
Middle ear: sound amplifier and reducer (impedance matching; protecting inner ear if too loud)
Inner ear: sound coder and transducer (turn vibration into neural action potentials (electric pulse)

4. Describe the responses in the basilar membrane to sounds of various frequencies.

From apical end to basal end, basilar membrane become progressively sensitive to higher frequency sounds.

5. What are hair cells? What role do they play in hearing?

Hair cells are the sensory receptors that perform the mechanical to electrical transformation. The electrical signals are carried by the auditory nerve to higher levels of the central nervous system for sound perception.

6. What is psychoacoustics?

Psychoacoustics is the branch of science that studies the relationships between physical properties of sound and the psychological correlates of these properties.

7. Describe Figures 5.9 and 5.10 on Pages 95 and 100 in your own words.

These two graphs describe how sound intensity is interpreted by the brain. Translation of dB SPL into loudness level uses 1000 Hz tone as a standard of comparison when estimating the levelof loudness of other frequencies. Notice that we are most sensitive to the mid-range frequency sounds. Also note that phon is the unit for loudness level. Sometimes we further translate equal loudness levels into a ratio scale of loudness called sones (2 sones is twice as loud as 1 sone).

8. What do the physical qualities of intensity and frequency correspond to in terms of subjective qualities?

Intensity => loudness
Frequency => pitch

9. What are the common scales for intensity, loudness, and pitch?

Intensity: watt/m^2 Intensity level: dB SPL (relative to hearing threshold for 1 kHz), dB IL (relative to hearing threshold for 1000 Hz), dB (reference is not the hearing threshold for1000 Hz tone)

Loudness: sone (1 sone = perceived loudness for 1000 Hz tone at 40 dB IL)
Loudness level: phon (1 phon is equivalent to the loudness level for 1000 Hz at 1 dB IL)
Pitch: mel
Frequency: Hz Note: Intensity and frequency are absolute physical quantities -- they are objective measurements of the sound properties. Loudness and pitch are perceptual quantities -- they are subjective measurements of the sound properties. Intensity level is a relative physical quantity based on a chosen reference. It equals 10*log (Ix/Ir), where Ix = the intensity of the sound, Ir = the intensity of the reference sound.Loudness level is the psychological correlate of intensity level. Its unit of measurement isphon. For the 1000 Hz tone, 1 dB IL corresponds to 1 phon (This is only true for the 1000 Hz tone and not for other frequencies).

10. What is meant by threshold?

Threshold refers to the intensity at which a sound can be detected 50% of the chance when presented to the listener. It is also known as minimum audibility, which essentially tests how sensitive the hearing system is to weak-energy sounds.

11. What are masking effects and binaural effects?

Masking effects refer to how the presence of one sound (masker) affects our detection of another (signal).
Binaural effects refer to how the brain interprets sound signals that arrive at both ears with differences in intensity and time. These effects are important for the judgment of sound localization and hemispheric lateralization.

12. True or False: 1 phon = 1sone = 1 dB SPL

F

13. True or False: 1 Hz = 1 mel

F

14. True or False: Otoacoustic emissions = tinnitus

F

15. True or False: differential threshold = difference limen = just noticeable difference

T

16. True or False: The middle ear always amplifies the incoming sound.

F