EE6424 Digital Audio Signal Processing
Part 2
Lecture 2:
Principal Characteristics of Speech

Outline of lecture

- Speech Communication
- Speech Production
  - Anatomy of vocal organs
  - Mechanism of speech production
- 语音感知[Speech Perception, EE6424 Part 01]
- Speech sounds
  - Pitch and formant
  - Voiced and unvoiced sounds
  - Vowel and consonant
  - Phonetic representation of speech
  - Source-filter model

EE6424 Part 2: Lecture 2.1

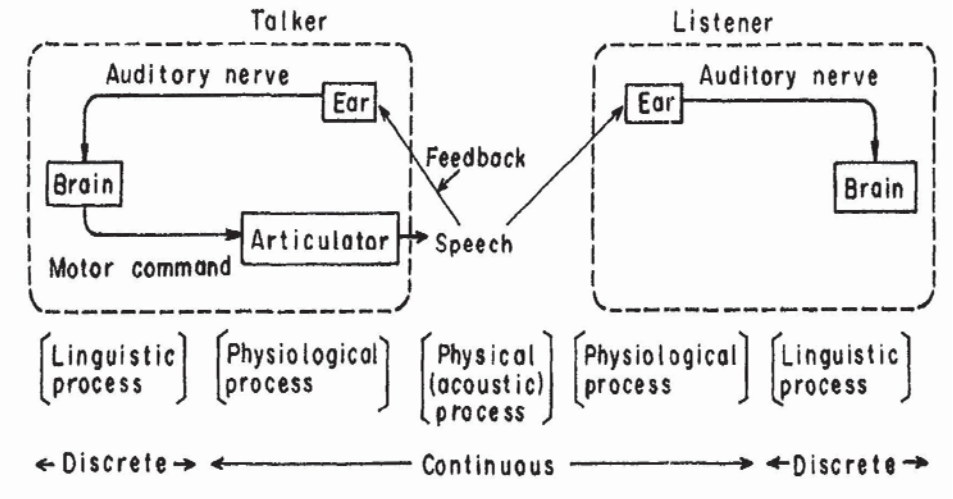
SPEECH COMMUNICATION

Speech chain

- Speech is used to communicate information from a speaker to a listener.
- Human speech production begins with an idea or thought that the speaker wants to convey to a listener.
- An acoustic sound pressure wave is produced through a series of neurological processes and muscular movements of vocal organs.
- The majority of the pressure wave originates from the mouth. Sound also emanates from the nostrils, throat, and cheeks.
- The acoustic wave received by a listener's auditory system is processed and converted back to neurological pulses.
- These pulses are interpreted in the auditory cortex of the brain to determine what idea was received.

内在联系

- The intrinsic connection (or interrelationship) between speech production and hearing is called the **speech chain**. It consists of the following stages:
  - Speech production (linguistic → physiological)
  - Acoustic process (transmission of acoustic sound pressure wave)
  - Speech perception (physiological → linguistic)
- The speech chain with its five stages (bottom) is shown below.



- The acoustic wave is transmitted back to the speaker's ears as well, providing necessary feedback for proper speech production.
- Any delay in this feedback to our own ears can cause difficulty in proper speech production.
- Loss of the feedback loop contributes significantly to the degradation in speech quality for individuals who have hearing disabilities.

cannot speak always cannot listen

Speech production and perception

- Speech is uttered for the purpose of being received and understood by the listener. Speech production is intrinsically related to hearing ability.
- Speech production process: 本质上
  - A speaker forms an idea and converts the idea into a linguistic structure by choosing appropriate words or phrases to represent the idea.
  - Order the words or phrases based on learned grammatical rules associated with the particular language.
  - Add additional local and global characteristics such as pitch intonation or stress to emphasize aspects important for overall meaning (e.g., a statement versus question).
  - The human brain then produces a sequence of motor commands that move various muscles of the vocal organs to produce the desired sound pressure wave.

- Speech perception process:
  - Begins when the listener collects the sound wave at the outer ear
  - Converts sound wave into neurological pulses at the middle and inner ear
  - Interprets the pulses in the auditory cortex to determine what idea was received.
- Refer to EE6424 Part 01 for more details on speech perception.

神经脉冲

Strength and limitation of vocal and hearing organs

- Speech production
  - Organs used in speech production are shared with other functions such as breathing, eating, smelling.
  - The multiple roles of these organs suggest that their present form may not be optimal for human communication.
  - Typical speech communication is limited to a bandwidth of 7 to 8 kHz.
- Speech Perception
  - Selectivity in what we wish to listen to. This permits the listener to hear to one individual voice in the presence of several simultaneous talkers, known as the cocktail party effects.
  - Inability to distinguish signals that are closely spaced in time or frequency (i.e., time and frequency masking)

Demo: cocktail party effects

- Source: http://youtube.com/watch?v=mN--nV61gDo

- http://vidque.com/303725/cocktail-party-effect

EE6424 Part 2: Lecture 2.2

THE PROCESS OF SPEECH PRODUCTION

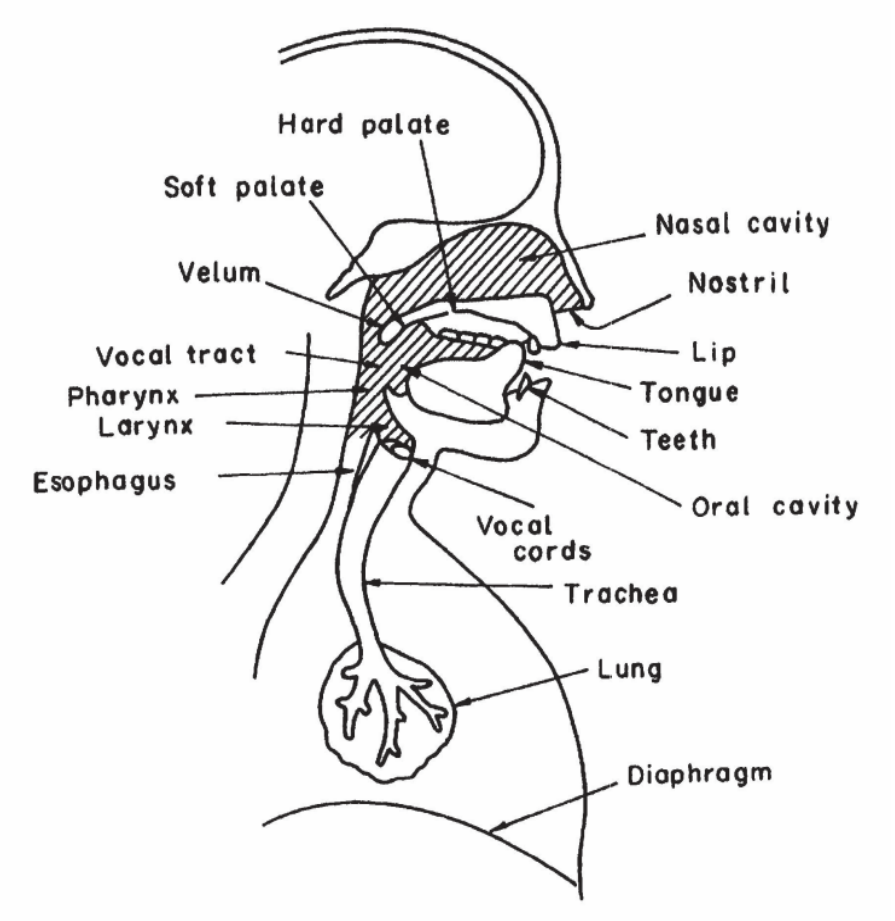
Anatomy of speech production system

喉咙

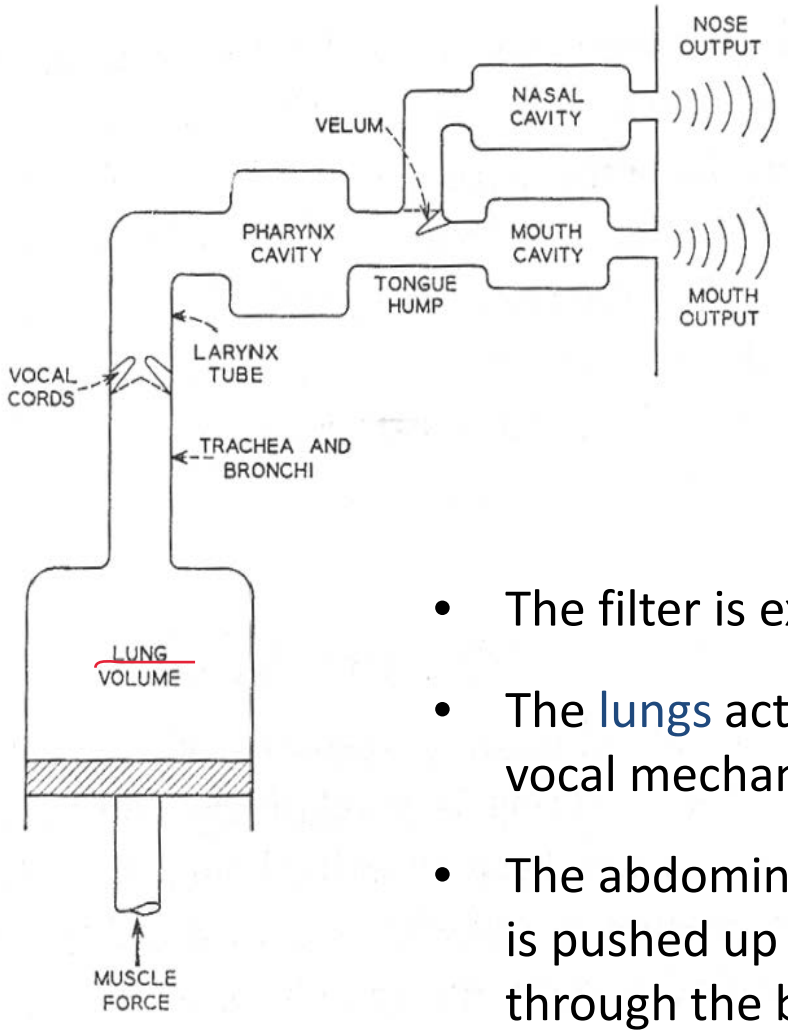
- Larynx – commonly known as the Adam's apple or voicebox. The larynx contains two small folds of muscle (or membrane) called the vocal folds (or vocal cords) which can be moved together or apart.
- Glottis – the space or opening between the two vocal folds, usually opens during breathing.
- Pharynx – the connection from esophagus to the mouth (i.e., the throat).
- Oral cavity – the mouth.
- Nasal cavity – often called nasal tract, begins at the velum (soft palate) and ends at the nostrils of the nose.
- Vocal tract – the cavity begins from the glottis to and ends at the lips. The vocal tract thus consists of the pharynx and oral cavity (i.e., the throat and month).

- The total length of the vocal tract is about 17 cm (14 cm) for adult male (female). The vocal tract length for an average child is 10 cm.
- The term vocal tract is often used in imprecise ways by engineers. It is used to refer to the combination of pharynx, oral cavity and nasal cavity. And more often to refer to the entire speech production system.
- Articulators – anatomical components including the vocal folds, velum, tongue, teeth, lips, and jaw that **move** to different positions to produce various speech sounds.
- The cross-sectional area of the vocal tract is determined by the positions of the tongue, lips, jaw, and velum. It varies from zero (complete closure) to about 20 cm^2.
- When the velum is lowered, the nasal cavity is coupled with the vocal tract to produce the nasal sounds of speech.

- Figure below shows a schematic diagram of human speech production system



Mechanism of speech production

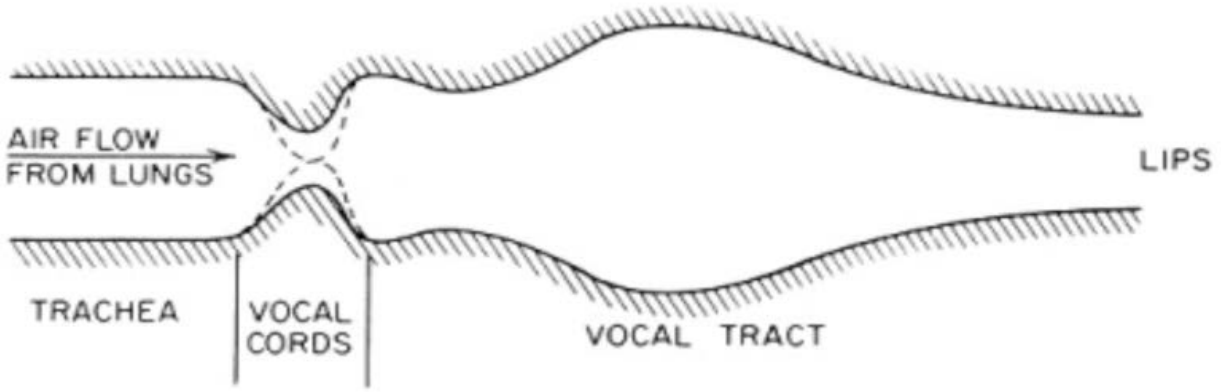


- The speech production process can be thought of as an acoustic filtering operation represented by a simplified acoustic model.
- The three main cavities (pharyngeal, oral and nasal cavities) comprise the main acoustic filter.
- The filter is excited by the organs below it.
- The lungs act as the source of air for exciting the vocal mechanism.
- The abdominal muscles force the diaphragm up, air is pushed up and out from the lung, passing through the bronchi, trachea and into the larynx.

- The articulators are used to change the properties of the system and the form of excitation
  - voiced sounds by the vocal folds vibration, while unvoiced sounds by constriction along the vocal tract.
  - Repositioning of the vocal tract articulators (tongue, lips, teeth, and jaw) causes the cross-sectional area of the vocal tract to change from zero to about 20 cm^2.
  - The nasal cavity constitutes an auxiliary path for the transmission of sound from the vocal tract, controlled by the size of the opening at the velum.
- The velum is lowered to produce the nasal sounds of speech.
- Adjusting the vocal tract shape to produce various linguistic sounds is called articulation.

Glottal airflow

- Figure below shows the side view of the vocal folds and the path for air flow from the lungs through the vocal tract.

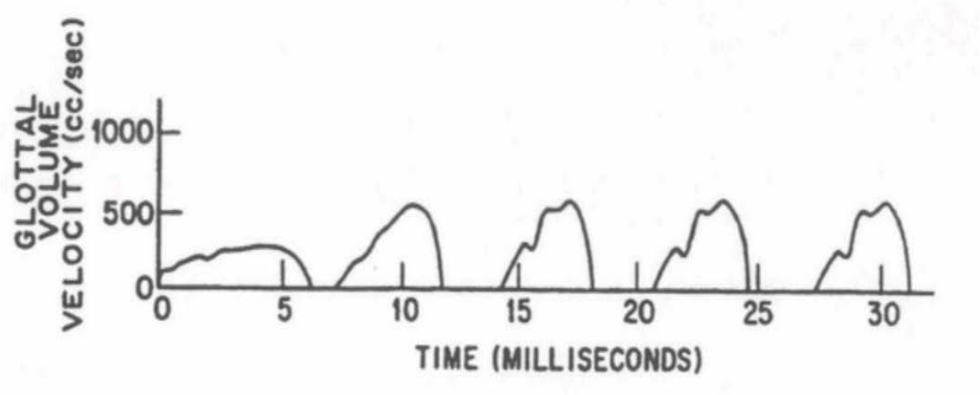


- To produce voiced sound, the two vocal folds become tensed via appropriate musculature control. Air from the lungs causes the tensed vocal folds to vibrate in the following manner:
  - Air pressure builds up behind the vocal folds and eventually blows them apart.
  - Air flows through the glottis and the air pressure drops allowing the vocal folds to close.
  - The cycle of building up pressure, blowing apart the vocal folds and then closing is repeated quasi-periodically as air continue to come out from the lungs.
- The periodic vibration of the vocal folds produces sound consisting of fundamental frequency and harmonics (i.e., not a pure sine wave).
- The positions of various vocal tract articulators determine the **tonal qualities** (i.e., the formants) of the sound that is produced.

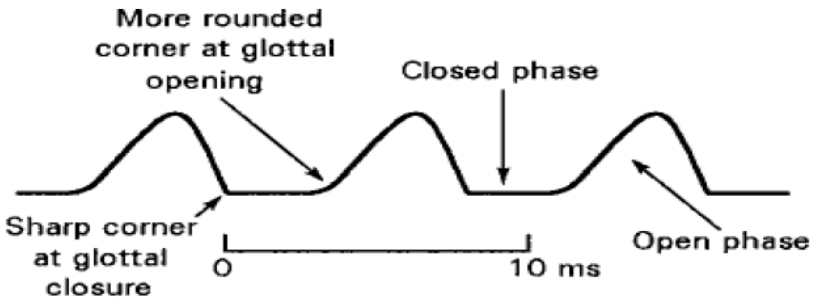
Glottal volume velocity

- The cycle of opening and closing of the vocal folds could be seen from the glottal volume velocity.

- The first 15 ms (one or two cycles) represents a period of buildup in the glottal flow until it begins to look like a quasi-periodic signal.



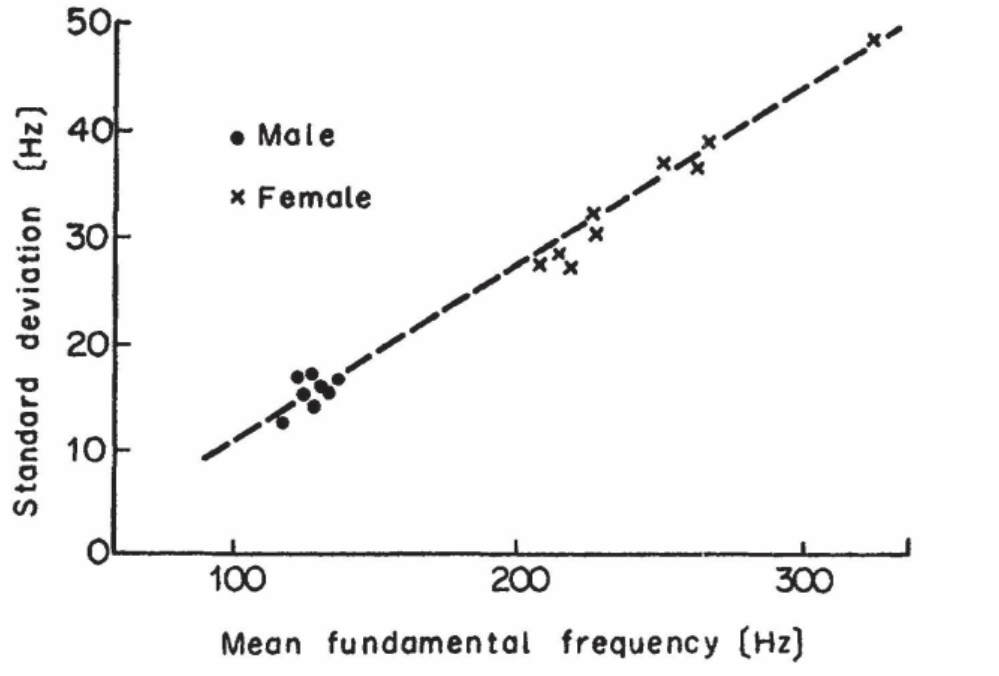
- The presence of sharp corner at the point of closure gives rise to harmonics of significant magnitude up to several kHz.



Pitch and fundamental frequency

- The vocal folds vibration period is called the **fundamental period** (or pitch period) T_0.
  倒数
- The reciprocal is called the **fundamental frequency** F_0 (or pitch), which corresponds to the rate of vibration.
- The average size of the vocal folds in men is larger than the average in women. The average fundamental frequency of an adult male in speaking a given utterance will often be lower than a female's.
- Each person has a pitch range to which he or she is constrained by his or her larynx:
  - Men: 50 – 250 Hz, Women: 120 – 500 Hz.
- In addition to pitch range, everyone has a habitual pitch level that is used naturally on the average.

- Statistical analysis of variation in fundamental frequency during conversation (measured for each speaker) indicates that the mean and standard deviation for female voices are roughly twice those for male voices. See figure below.



Pitch and intonation

- Accent and intonation are results of temporal variation of the fundamental period (above and below a person habitual pitch level).
- When the vocal folds are strongly strained and the pressure of the air rising from the lungs is high, the vocal folds vibration period becomes short and the pitch of the sound source becomes high.
- A lower-pitched sound is produced by lowering the sub-glottal air pressure and a weaker strain on the vocal folds.
- Pitch is shifted up and down in speaking in response to intonation, stress, and emotion. [See Deller, pp. 114]

Formants 共振

- As sound propagates down the vocal tract, the frequency spectrum is shaped by the frequency selectivity of the cavities with well-defined regions of emphasis (resonances) and de-emphasis (anti-resonances).
- The resonance frequencies of the vocal tract tube are called formant frequencies, or simply, formants. More precisely, formant frequencies refer to the nominal center frequencies of the resonances.
- The formant frequencies depend upon the shape and dimensions of the vocal tract. Conversely, each shape is characterized by a **set** of formant frequencies.
- Different sounds are formed by varying the shape of the vocal tract (声道) (or more precisely, by re-positioning the articulators). The spectral properties of the speech signal vary with time as vocal tract shape varies.

- This results in a speech signal as sequence of sounds that carries the information from a speaker to a listener.
- The term formant was used to reflect the frequency selectivity of the resonances which “form” the overall spectrum.

Physical constraints of larynx

near periodical

- The oscillation of the vocal folds during voiced speech is quasi-periodic as cycle-to-cycle variation could be observed as two types of measures
  - Jitter (frequency perturbation)
  - Shimmer (amplitude perturbation)
- In sustained phonation of normal voice, the jitter is about 1% in frequency and the shimmer is about 6% in amplitude.

Artificial larynx

- The larynx plays a significant role in speech production.
- An artificial larynx is used as an aid to patients who had their larynx removed by operation.
- It contains a vibrating diaphragm that produces **quasi-periodic sound** that can be coupled directly into human vocal tract by holding the device against the neck.
- By using the “on-off” and “rate-of-vibration” control an experienced human user can create an appropriate excitation signal that mimics the one produced by the vibrating vocal folds, thereby enabling the user to create speech for communication with other humans.

Video: TruTone

- Source: YouTube http://www.youtube.com/watch?v=AYydnhu6NbU

EE6424 Part2: Lecture 2.3

SPEECH SOUNDS

Voiced and unvoiced speech sounds

- Speech sounds are either voiced or unvoiced. The speech sounds generated **with** vocal folds vibration are referred to voiced sounds and those **without** are named unvoiced sounds.
- Experiment: consider pronouncing the syllable “fa” as in the word “father” while placing your finger on the front of your neck.
  - Pronounce the /f/ sound alone for a few seconds. Next, pronounce the /a:/ sound alone for a few seconds.
  - Vibration in the front your neck could only be felt in the later case. Speak louder if you have problems feeling it!
- Voiced sounds are produced by forcing air through the glottis. The glottis becomes narrower and the tension of the vocal folds is adjusted so that they vibrate.

- Unvoiced sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction. During the process, the vocal folds are relaxed (non-tense) and spread apart.
- In addition to vocal folds vibration, two other mechanisms are responsible for changing the airflow from the lung into speech sounds:
  - Noise like sounds produced by turbulent flow which occurs when the airflow passes through a constriction in the vocal tract by the tongue or lips. **Fricatives**, such as, /s/, /f/, and /ʃ/.
  - Impulsive sounds which occur with the sudden release of high-pressure air by using the tongue or lips. **Plosives** (stop consonants), such as, /p/, /t/, and /k/.
- There are fricative and plosive which are generated in conjunction with vocal folds vibration, whereby two-types of excitations (vibration and constriction) co-exist. These are referred to as voiced consonants.

Phonetic representation of speech

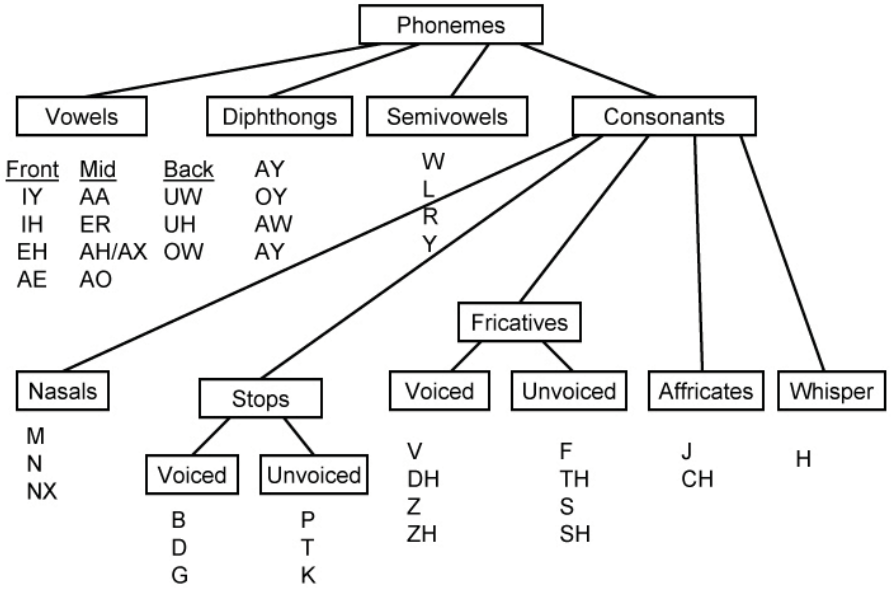
- Words are pronounced in terms of individual speech units called phones. Pronunciation of every word is described by phonetic alphabets. For example, the pronunciation of the word “week” is /wi:k/.
- A phone is a speech sound. Speech signals are composed of sequence of phones that serve as the symbolic representation for a thought that the speaker wish to relay to the listener.
- Vocal tract articulators consists of human tissue, their position from one phone to the next is executed by movement of muscles.
- Speech is not a precisely a string of discrete well-defined phones, but rather a series of steady state of target sounds with intermediate transition.
- The arrangement of phones in an utterance is governed by rules associated with a language.

Vowels and consonants

辅音 元音

- Phones are divided into two main classes: consonants and vowels. Both kinds of sounds are formed by the motion of air through the mouth, throat or nose.
- Vowels are generally voiced, and are louder and longer lasting than consonants. Diphthongs are vowel combinations.
- Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced (i.e., with or without vocal folds vibration).
- Consonants which are accompanied by vocal folds vibration are called voiced consonants, and those which are accompanied by this vibration are called unvoiced consonants.

- For American English, there are between 39 and 48 phones. Figure below shows a reduced set of 39 phones:



- 11 vowels
- 4 diphthongs (vowel combinations)
- 4 semi-vowels (vowel-like consonants)
- 3 nasal consonants
- 6 stop/plosive consonants (voiced and unvoiced)
- 8 fricative consonants (voiced and unvoiced)
- 2 affricate consonants
- 1 whispered sound (/h/ as in “hat”)

- Semivowel, nasal, and affricates are considered as consonants:
  - Semivowels are vowel-like consonants
  - Affricates are produced by the succession of plosive (stop) and fricative sounds
  - Nasal consonants are produced by coupling airflow to the nasal cavity when the velum is lowered.
  - Whispering is produced when a turbulent flow of air is made at the glottis by slightly opening the vocal folds so that vocal folds vibration is not produced.

CMU pronunciation dictionary

- The table shows a set of 39 phones used in the CMU pronunciation dictionary
- Available online at http://www.speech.cs.cmu.edu/cgi-bin/cmudict (different being the whispered sound /h/ being considered as fricatives in the table)

Class	ARPAbet	Example	Transcription
Vowels and diphthongs	IY	*b*e*e*t	[B IY T]
	IH	*b*i*t*	[B IH T]
	EY	*b*a*i*t	[B EY T]
	EH	*b*e*t*	[B EH T]
	AE	*b*a*t*	[B AE T]
	AA	*b*o*b*	[B AA B]
	AO	*b*o*r*n	[B AO R N]
	UH	*b*oo*k*	[B UH K]
	OW	*b*oa*t*	[B OW T]
	UW	*b*oo*t*	[B UW T]
	AH	*b*u*t*	[B AH T]
	ER	*b*i*rd*	[B ER D]
	AY	*b*u*y*	[B AY]
	AW	*d*o*w*n	[D AW N]
	OY	*b*o*y*	[B OY]
Glides	Y	*y*ou	[Y UH]
	R	*r*ent	[R EH N T]
Liquids	W	*w*it	[W IH T]
	L	*l*et	[L EH T]
Nasals	M	*m*et	[M EH T]
	N	*n*et	[N EH T]
	NG	si*ng*	[S IH NG]
Stops	P	*p*at	[P AE T]
	B	*b*et	[B EH T]
	T	*t*en	[T EH N]
	D	*d*ebt	[D EH T]
	K	*k*it	[K IH T]
	G	*g*et	[G EH T]
Fricatives	HH	*h*at	[HH AE T]
	F	*f*at	[F AE T]
	V	*v*at	[V AE T]
	TH	*th*ing	[TH IH NG]
	DH	*th*at	[DH AE T]
	S	*s*at	[S AE T]
	Z	*z*oo	[Z UW]
	SH	*sh*ut	[SH AH T]
	ZH	a*z*ure	[AE ZH ER]
Affricates	CH	*ch*ase	[CH EY S]
	JH	*j*udge	[JH AH JH]

The world languages

- There are about 5000 to 8000 languages in the worlds, each with its own particular selection of phones.

Each dot represents the geographic center of the 6,909 living languages in the world. [Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.]

- The human speech production and perception mechanisms are identical across human races.
- Languages make their selection from the set of humanly possible sounds. Therefore, the inventory of phones overlaps significantly among languages.
- The most unusual sound in English is the phoneme /th/ as in the words “think” and “this”.
- The total number of phones requires to represent all the sounds of the world languages ranges from 200 to 300.
- This idea was used in constructing the International Phonetic Alphabet (IPA). Phones that are common in languages are grouped together and given the same symbol in IPA.

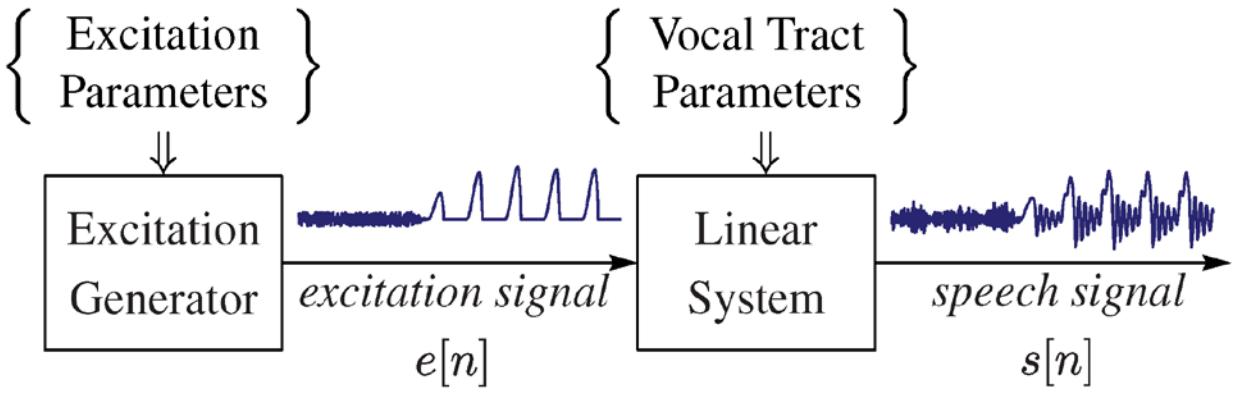
Phoneme versus phone

- The basic unit for describing how speech conveys linguistic meaning is called a phoneme.
- Each phoneme can be considered as a code that consists of a unique set of articulator gestures (type and location of sound excitation as well as the location of the articulators).
- The actual sounds that are produced in speaking are called phones. A phone is the physical speech sound that conveys a phoneme.
- When spoken in context, there is a period of transition between phonemes, which slightly modified the manner a phoneme is produced.
- The interplay between phonemes produced in an utterance is called co-articulation.

Source-filter model

- Voiced sounds are produced when the vocal tract is excited by pulses of air pressure resulting from quasi-periodic opening and closing of the glottis.
- Unvoiced sounds are produced by forcing air through a constriction formed (or by momentarily closing and releasing of airflow) creating a random noise excitation.
- Both quasi-periodic and random noise excitation sources create a wide-band signal to the vocal tract with certain vocal tract shape-dependent resonances that tends to shape the spectrum of the excitations.
- The fine structure of the time waveform is created by the sound source, and the frequency response of the vocal tract shapes these sounds into phones.

- A source-filter model models speech as the output of a slowly time-varying digital filter with an excitation that captures the nature of voiced/unvoiced distinction in speech production.



- A source-filter model consists of two components:
  - An excitation generator simulating different modes of the sound generation
  - A discrete-time time varying filter simulates the frequency response of the vocal tract
- The assumptions are as follows:
  - The source and the filter does not interact
  - Linearity, i.e., $T(a+b) = T(a) + T(b)$
  - The changes in vocal tract shape occur relatively slowly. The linear system could be fixed over time interval of 10 ms or so.

- A general form of the filter include poles (c_k) and zeros (d_k). Many applications only include poles in the model for simplicity.

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=0}^{N} a_k z^{-k}} = \frac{b_0 \prod_{k=1}^{M} (1 - d_k z^{-1})}{\prod_{k=1}^{N} (1 - c_k z^{-1})}$$

- For unvoiced speech, the excitation is a random noise with a flat spectrum.
- For voiced speech, the excitation is a sequence of periodic pulses with a harmonic line spectrum. The period of the pulses determined the perceived pitch.

Summary

- Speech processing relies on the basic research on speech and hearing science
- Speech processing engineers need to understand the basic concepts from these areas in order to analyze and model speech for applications like speech recognition, speaker recognition and speech synthesis.
- Speech sounds could be voiced or unvoiced:
  - Voiced sounds are produced with vocal folds vibration (pitch period, fundamental frequency, glottal volume velocity)
  - Unvoiced sounds are produced by constriction (noise like turbulent or impulsive)
  - Consonants could be voiced or unvoiced. Vowels are mostly voiced.
- Given the excitation, the frequency response of the vocal tract shapes sounds into phones.