Speech Communication, Spring 2006

- Intelligent Multimedia Program -

Lecture 1: Introduction, Speech Production and Phonetics

Zheng-Hua Tan
Speech and Multimedia Communication Division
Department of Communication Technology
Aalborg University, Denmark
zt@kom.aau.dk

Part I: Introduction

- Introduction
  - Problem definition
  - State-of-the-art
  - Course overview
- Speech production and acoustic phonetics
  - The anatomy of speech production
  - Articulatory phonetics
  - Acoustic phonetics
  - Models of speech production

Computer as dream of human being

- HAL talks, listens, reads lips and solves problems
  - Nature and effortless for huamn
  - Hard for computer
  - Dream of AI scientists and human
  - True in *2001: A Space Odyssey*

(*After 2001: A Space Odyssey, 1968*)

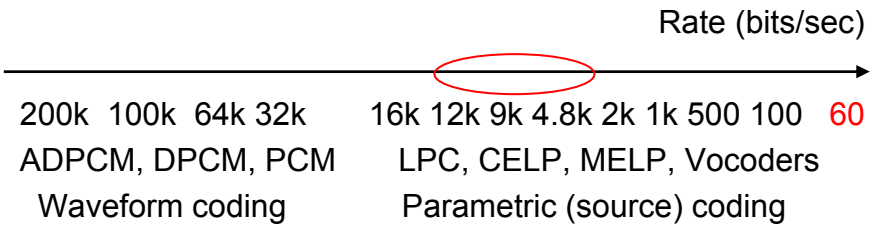
Computer as a reality: state-of-the-art

- Demo
  - Microsoft demo video

  - Text to speech (TTS)
    - Festival TTS @ CSTR Edinburg University
    - Next generation TTS @ AT&T

Information in Speech

- Speech coding data rates



Human can understand text:

10 char/sec x 6 bits/ASCII char = 60 bits/sec

Is content in speech more than 60 bits/sec?

Information in Speech – cont.

- Examples

"That's one **small step for man**; one **giant leap for mankind.**"

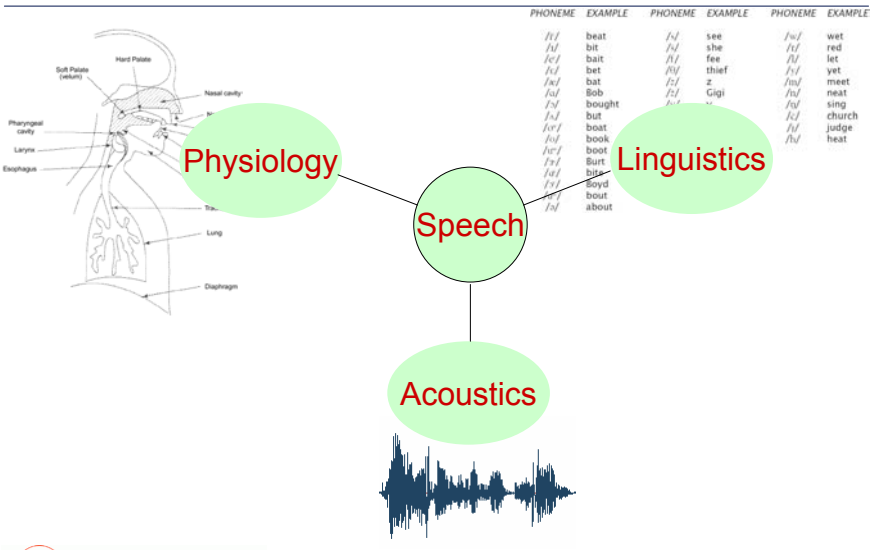
-- Neil Armstrong, *Apollo 11 Moon Landing Speech*

"**I have a dream** that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character. I have a dream today!"

-- Martin Luther King, Jr., I Have a Dream

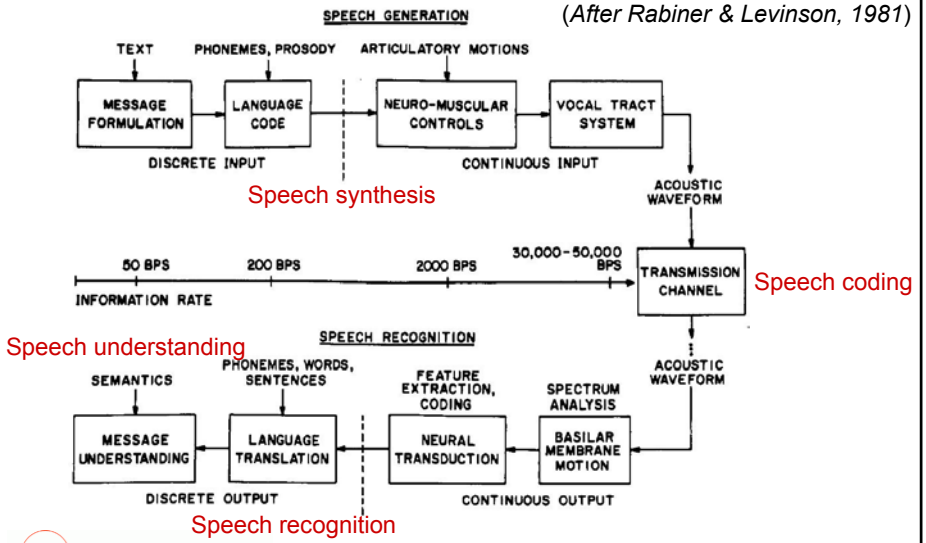
Speech contains speaker identity, emotion, meaning, text. → speech techniques

Speech is a complex process



Human speech communication process

(*After Rabiner & Levinson, 1981*)



Study topics and applications

- Introduction
- Speech Production and Acoustics Phonetics
- Speech Analysis and Speech Synthesis
- Speech Coding
- Speech Recognition
- Speech-Related Tools and Applications

Course Outline

- MM1 – Speech production, acoustic phonetics and speech modelling
  - The anatomy of speech production
  - Phonetics
  - Models of speech production
- MM2 – speech analysis
  - Speech perception and its models
  - Short-term processing of speech
  - Linear prediction analysis
  - Cepstral analysis
- MM3 – speech coding and synthesis
  - Speech synthesis
  - Speech coding
- MM4 - speech recognition
  - Introduction
  - DTW based speech recognition
  - HMM
- MM5 – speech recognition
  - HMM based speech recognition
  - HTK, token passing

Literature

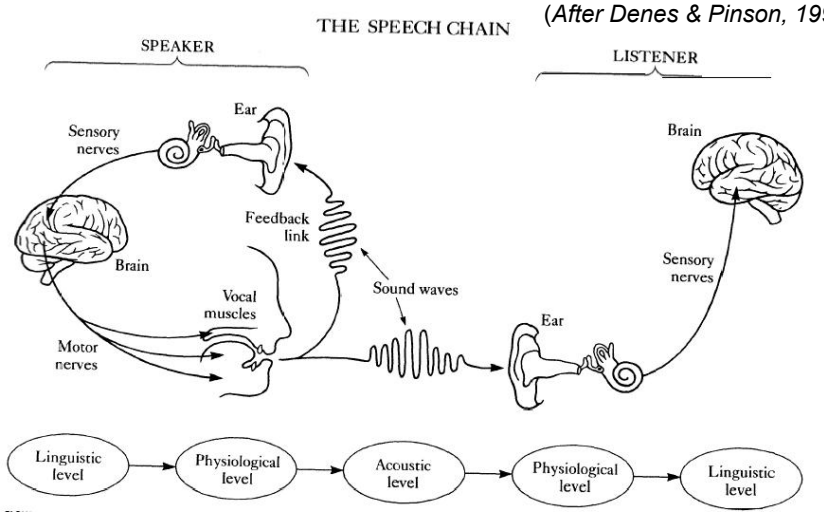
- Textbook:
  - J Deller, J Hansen and J Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, 2000.
- Reading:
  - Huang, Acero and Hon, Spoken Language Processing, Prentice-Hall, 2001.
  - D. O'Shaughnessy, Speech Communications, IEEE Press, 2000
  - Rabiner and Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.
  - Rabiner and Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.

Part II: Speech production

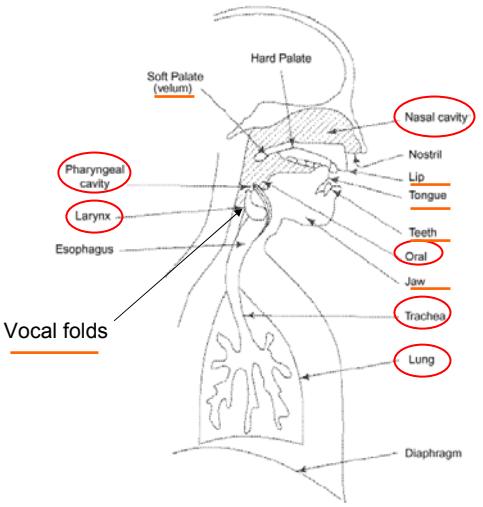
- Introduction
- Speech production, acoustic phonetics and speech modelling
  - The anatomy of speech production
  - Articulatory phonetics
  - Acoustic phonetics
  - Models of speech production

The speech chain

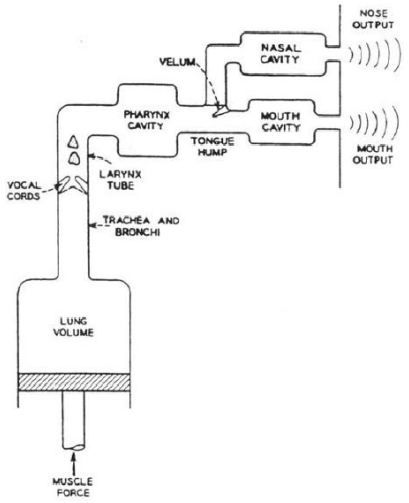
(*After Denes & Pinson, 1993*)



Schematic diagram of speech production

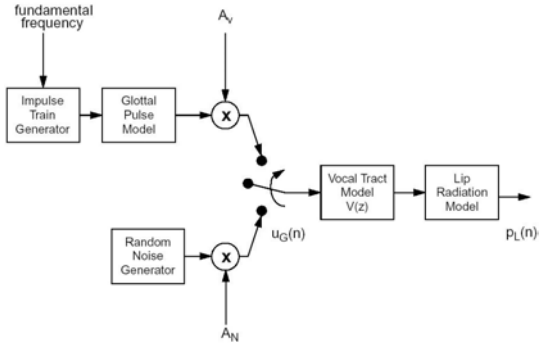


Block diagram of speech production



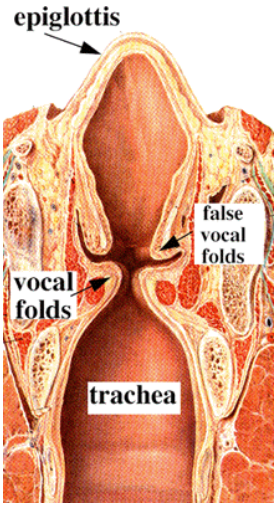
Model of speech production

- Digital model of speech production



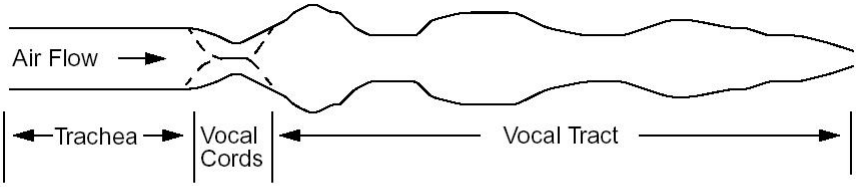
Cross section of the larynx

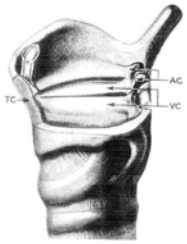
- Larynx: the source of most speech
- Vocal cords (folds): the two folds of tissue in the larynx. They can open and shut like a pair of fans.
- Glottis: the gap between the vocal cords. As air is forced through the glottis the vocal cords will start to vibrate and modulate the air flow.
- This process is known as phonation.
- The frequency of vibration determines the pitch of the voice (for a male, 50-200Hz; for a female, up to 500Hz).



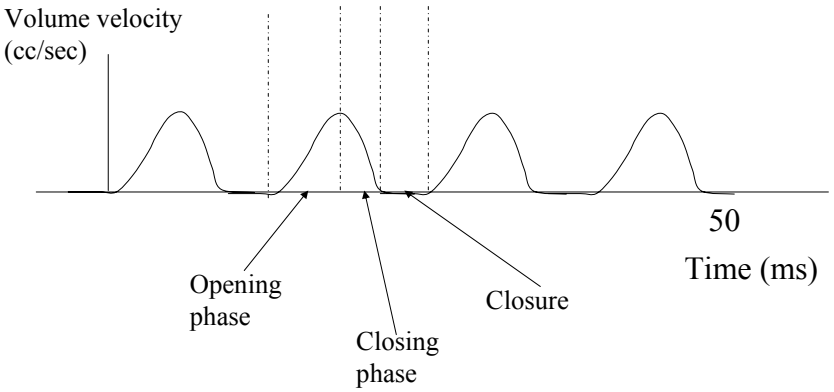
Vocal cords

- Vocal cords form a relaxation oscillator (voiced excitation)





Glottal flow

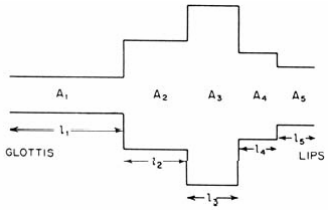


Pitch Period = 12.5ms
Fundamental frequency = 1/.0125 = 80Hz

Vocal tract modelling

- Source-filter model

- Vocal tract is a concatenation of tubes with varying cross-sectional areas



Type of excitation

- Voiced: produced by forcing air through the glottis
  - vowels (inc. diphthongs) are voiced
- Unvoiced: generated by forming a constriction at some point along the vocal tract and forcing air through the constriction

Role of the vocal tract

- Vowels: produced by exciting a fixed vocal tract with quasi-periodic pulsed of air caused by vibration of the vocal cords
- Consonants: a significant restriction and thus weaker in amplitude and noisy-like
- Formants: resonances determined by the shape of vocal tract, which form the overall spectrum and the properties of the filter

The speech signal

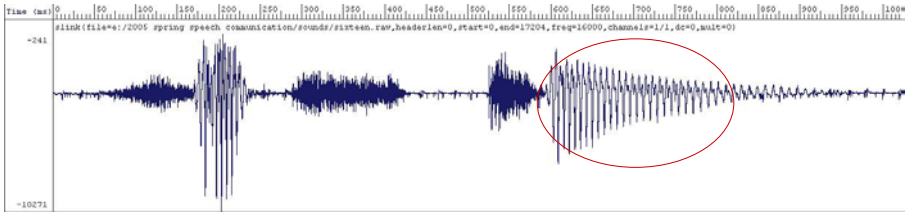
- Speech is a sequence of highly changing sounds
- When producing sounds, the vocal cords and the various articulators slowly change over time
- There is a need to study speech sounds, their production, and the signs used to represent them → phonetics

Phonetics

- Phonetics: study of speech sounds, their production, and the signs used to represent them.
  - articulatory phonetics: how they are made by moving various organs in the vocal tract.
  - acoustic phonetics: how they are perceived by the human ear and their physical properties. The study is conducted by observing and measuring the speech waveform and spectrum.

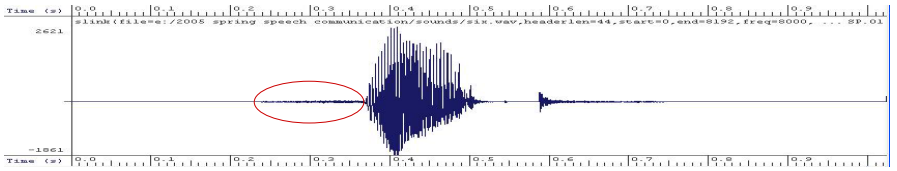
Speech sounds and waveforms

sixteen /s/ /i/ /k/ /s/ /t/ /ee/ /n/

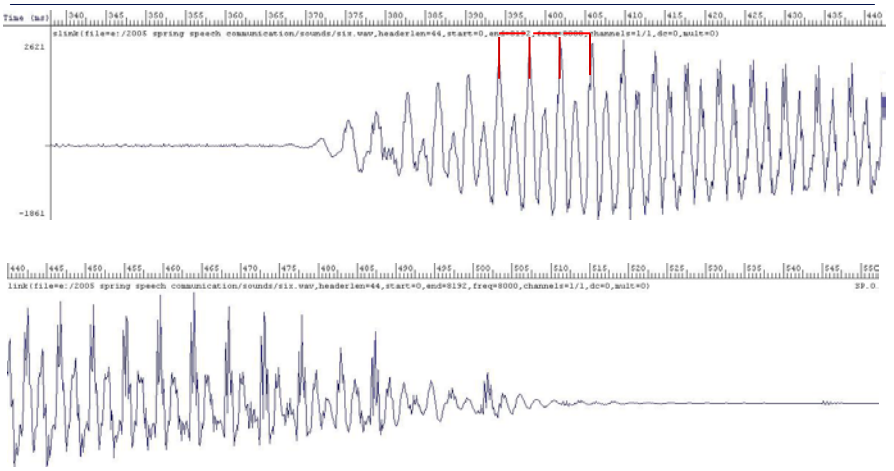


six

periodicity, intensity, duration, boundary, etc



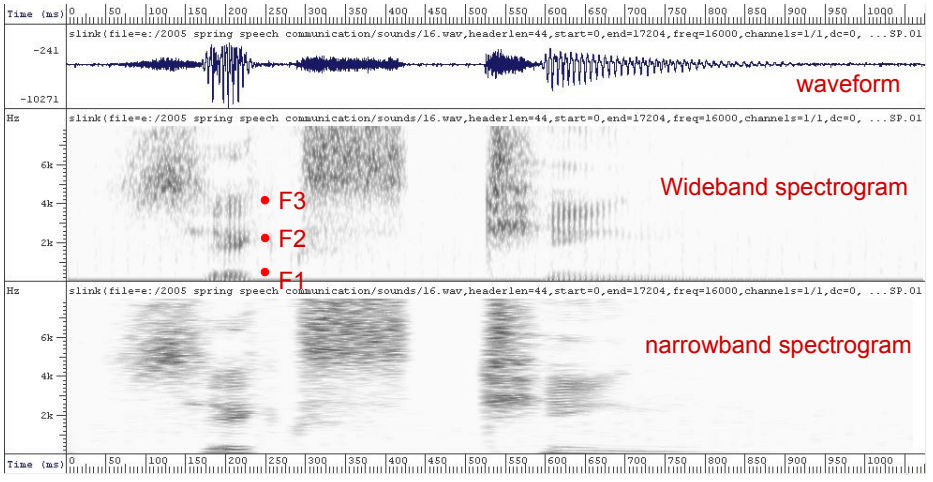
Observing pitch from waveforms



Spectrogram

- Spectrogram
  - two-dimensional waveform (amplitude/time) is converted into a three-dimensional pattern (amplitude/frequency/time)
  - Wideband spectrogram: analyzed on 15ms sections of waveform with a step of 1ms
    - voiced regions with vertical striations due to the periodicity of the time waveform (each vertical line represents a pulse of vocal folds) while unvoiced regions are solid/random, or 'snowy'
  - Narrowband spectrogram: on 50ms
    - pitch for voiced intervals in horizontal lines

Sound Spectrogram: an example



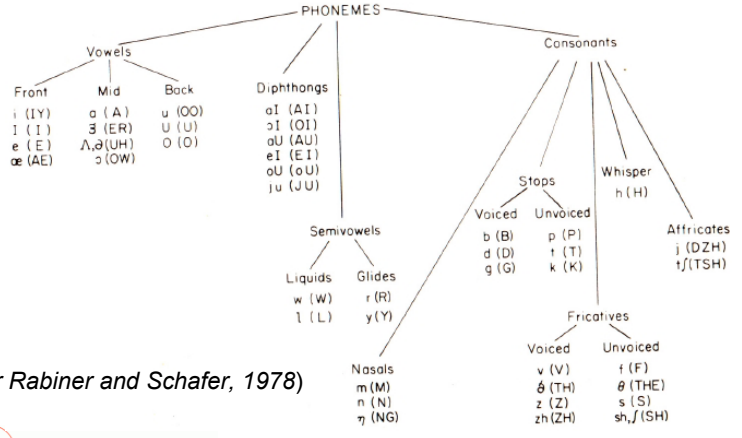
Phonemes in American English

	VOWELS:	DIPHTHONGS:	FRICATIVES:	NASALS:
Front	/i/ heed	/Y/ hide	/v/ van	/m/ mom
	/I/ hid	/W/ how'd	/D/ then	/n/ noon
	/e/ hayed	/O/ boy	/z/ zebra	/G/ sing
	/E/ head	/X/ rose	/Z/ measure	
	/@/ had		/f/ fan	**STOPS:**
	/R/ heard	**SEMI-VOWELS:**	/T/ think	/b/ bag
	/x/ ago	**Liquids**	/s/ sit	/d/ dog
	/A/ mud	/r/ ran	/S/ shoe	/g/ goat
	/u/ who'd	/l/ liquid	/h/ help	/p/ peal
Back	/U/ hood	**Glides**		/t/ tea
	/o/ hoed	/w/ want	**AFFRICATES:**	/k/ kick
	/c/ hawed	/y/ yard	/J/ just	
	/a/ hod		/C/ channel	

(*After J. Hansen*)

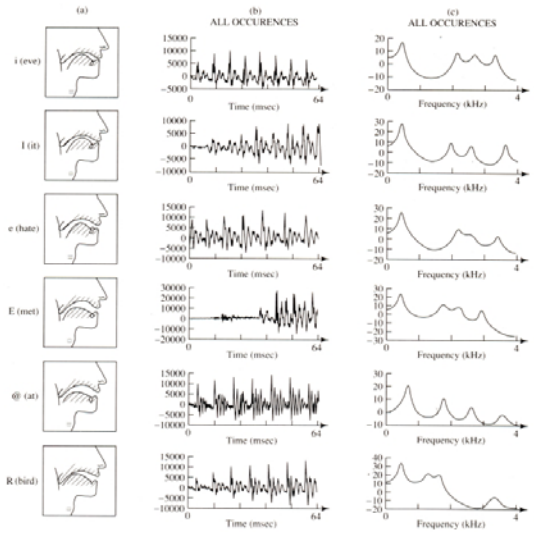
Phoneme classification chart

- Sound categorization according to the position of the articulators.



(*After Rabiner and Schafer, 1978*)

Vowel production: examples

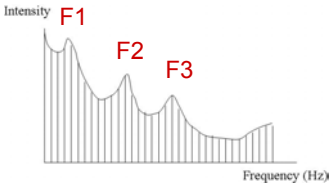


(*After Joseph Picone*)

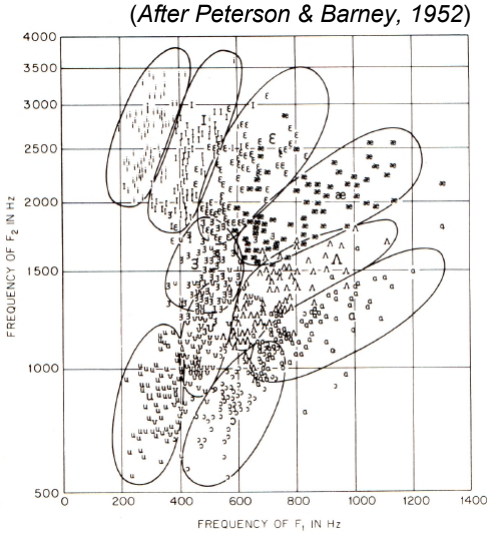
- Fixed vocal tract shape
- Voiced
- Cross-sectional area → F_i
- Tongue position
- → sound

The vowel space

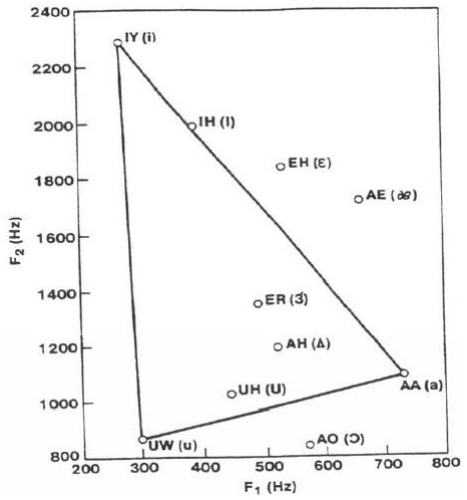
- by the locations of the first and second formant frequencies:



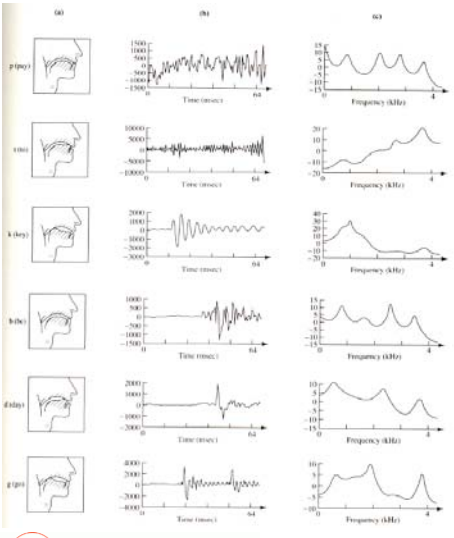
(*After Peterson & Barney, 1952*)



The vowel triangle



Consonant production: examples



(*After Joseph Picone*)

Diphthongs

- A diphthongs involves an intentional movement from one vowel toward another vowel
- Differ from two distinct vowels: representing a transition from one vowel target to another, yet neither vowel is actually reached
- Diphthongs: (Fig. 2.14, pp129, John3 2000)
  - /Y/ hide
  - /W/ down
  - /O/ boy
  - /X/ rose

Semivowels

- Vowel-like, but weaker than most vowels due to their more constricted vocal tract
- Voiced
- Semivowels: (Fig. 2.15, pp130, John3 2000)
  - Liquids:
    - /r/ ran
    - /l/ liquid
  - Glides:
    - /w/ want
    - /y/ yard

Nasals

- Produced by the glottal waveform exciting an open nasal cavity and closed oral cavity.
- Similar to vowel but weaker due to limited ability of the nasal cavity to radiate sound
- Nasals:
  - /m/ moon
  - /n/ noon
  - /G/ sing

Fricatives

- Produced by exciting the vocal tract with a steady air-stream that becomes turbulent at some point of constriction
- Fricatives

/v/ van
/D/ then
/z/ zebra
/Z/ measure
/f/ fan
/T/ think
/s/ sit
/S/ shoe
/h/ help

Affricates

- formed by transitions from a stop to a fricative
- Affricates:
  - /J/ just
  - /C/ channel

Stops (or Plosives)

- Stops consonants are transient, non-continuant sounds that are produced by building up pressure behind a total constriction somewhere along the vocal tract, and suddenly releasing this pressure
- Stops

/b/ bag
/d/ dog
/g/ goat
/p/ peal
/t/ tea
/k/ kick

Speech Tool

- **Speech Filing System- Tools for Speech Research**
  - It performs standard operations such as recording, replay, waveform editing and labelling, spectrographic and formant analysis and fundamental frequency estimation.
  - **http://www.phon.ucl.ac.uk/resource/sfs/**

Summary

- Speech technology
- The speech chain
- Anatomy of speech production
- Speech signals: waveform and spectrogram
- Phonetics
- Modelling

- Next lecture: Speech Analysis