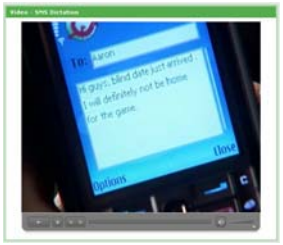
Multi-Modal User Interaction
Fall 2008

Lecture 1: Introduction

Zheng-Hua Tan

Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk

Aperitif



About the course

- **Purpose**
  - To give the student a comprehension of the principles for multi-modal interaction, in particular speech-based interfaces
  - To enable the student to extend the methods for HCI GUI design to analyse, design and synthesise multi modal user interfaces
- **Contents**
  - Automatic speech recognition and –synthesis
  - Integration of information from e.g. speech and visual modalities into advanced multimodal interfaces
  - Architectures and platforms of MM systems
  - Multi modal interface design and evaluation methods

Course Outline

- MM1~5: Speech synthesis and recognition
  - Introduction
  - Speech synthesis
  - Speech recognition
- MM6 ~10: Multimodal interaction
  - Integration of information from multiple modalities
  - Architectures and platforms of MM systems
  - Multi modal interface design and evaluation methods

Literature

- Textbook:
  - McTear, Spoken Dialogue Technology, Springer, 2004.
- Readings:
  - 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.

Course homepage and contact info

- http://kom.aau.dk/~zt/cources/MMUI/
- Zheng-Hua Tan
  - +45 9940-8686
  - Office: Room A6-319, Niels Jernes Vej 12
- Lars Bo Larsen
  - +45 9940-7202
  - Office: Room A6-317, Niels Jernes Vej 12

Part I: Introduction

- Introduction
  - Speech input and output – components of speech interaction
  - State-of-the-art
- Basics about speech – a short introduction
- Speech synthesis

Computer as dream of human being

HAL talks, listens, reads lips and solves problems

- Nature and effortless for human
- 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

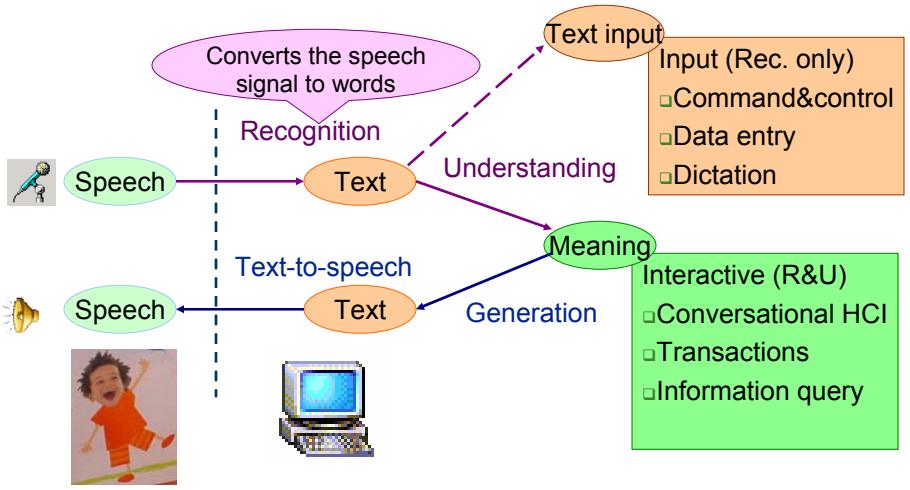
- Man vs. machine ➡

- Text to speech (TTS)
  - Next generation TTS @ AT&T

State-of-the-art

- Dragon Naturally speaking 10
  - It's three times faster than most people type
  - Up to 99% accurate right out of the box!
  - The latency between speaking and seeing words on the PC has nearly been eliminated.
  - Let you find files on your PC, search web maps, shop on eBay, set appointments and more, all with simple voice commands.

Human-computer interaction via speech

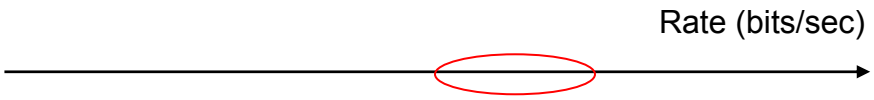


Part II: Basics about speech

- Introduction
- Basics about speech – a short introduction
- Speech synthesis

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?

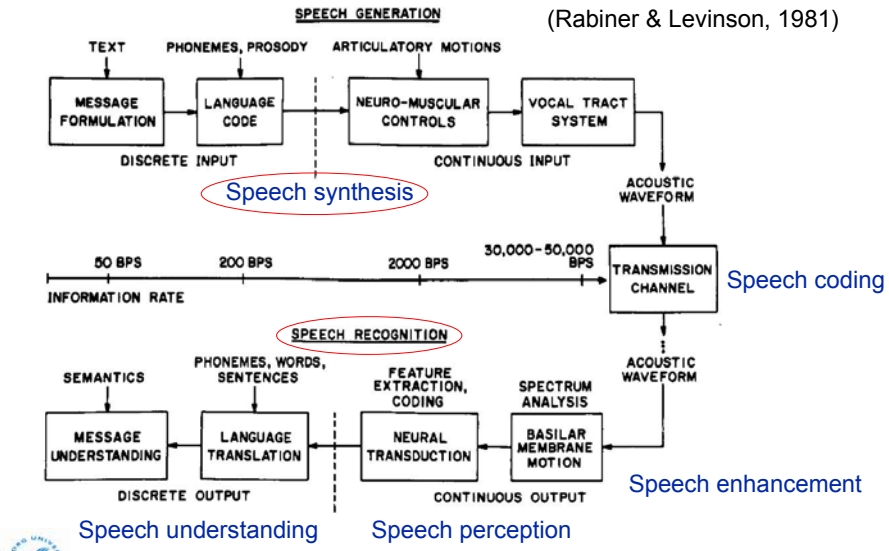
Information in Speech – cont.

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

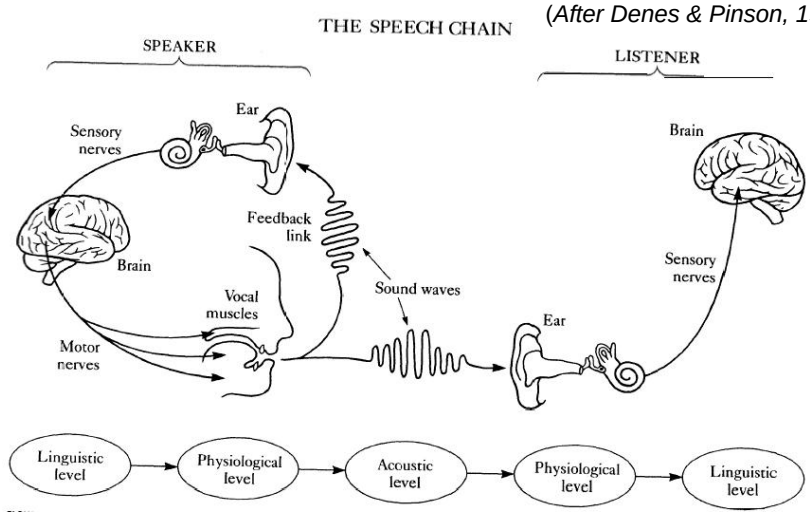
-- Neil Armstrong, *Apollo 11 Moon Landing Speech*

Speech contains speaker identity, emotion, meaning, text, language, sex and age, channel characteristics. → speech techniques

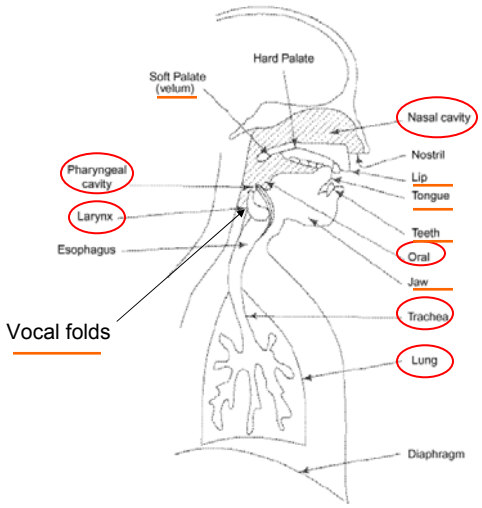
Human speech communication process



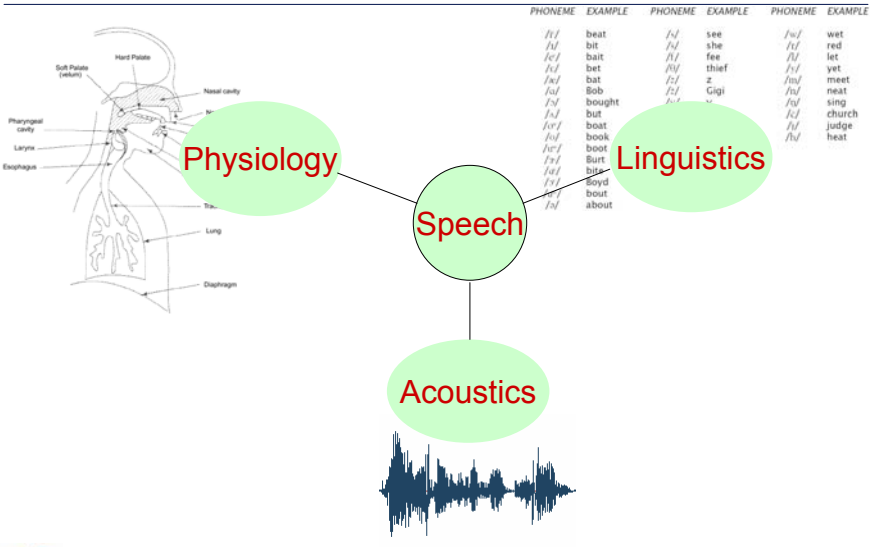
The speech chain



Schematic diagram of speech production

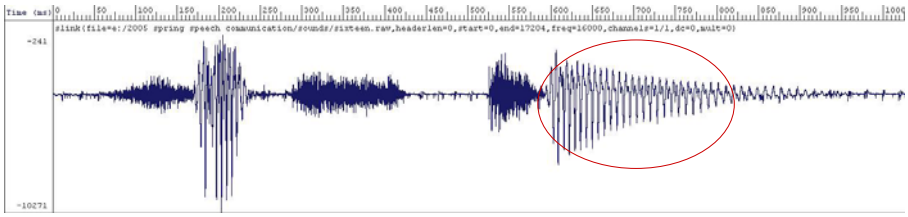


Speech is a complex process

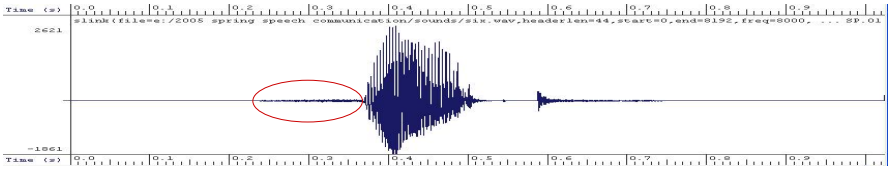


Speech sounds and waveforms

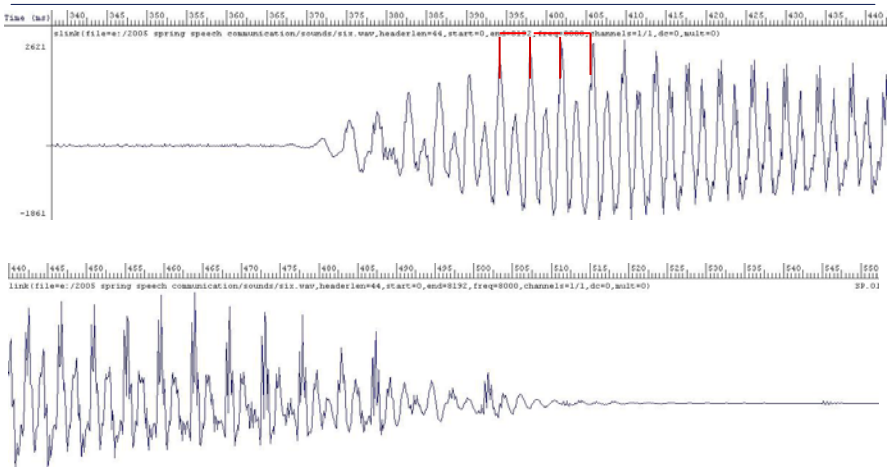
sixteen /s/ /i/ /k/ /s/ /t/ /ee/ /n/



six periodicity, intensity, duration, boundary, etc



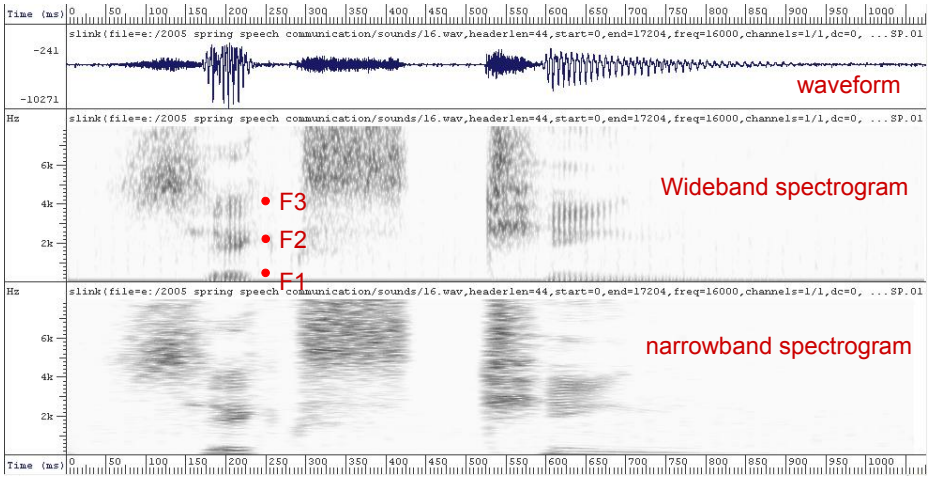
Observing pitch from waveforms



Spectrogram

- Spectrogram
  - 2-D waveform (amplitude/time) is converted into a 3-D 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 'snowy'.
  - Narrowband spectrogram: analyzed on 50ms sections
    - Pitch for voiced intervals in horizontal lines

Sound Spectrogram: an example



Phonemes in American English

VOWELS:

Front
- /i/ heed
- /I/ hid
- /e/ hayed
- /E/ head
- /@/ had

Back
- /R/ heard
- /x/ ago
- /A/ mud
- /u/ who'd
- /U/ hood
- /o/ hoed
- /c/ hawed
- /a/ hod

DIPHTHONGS:
- /Y/ hide
- /W/ how'd
- /O/ boy
- /X/ rose

SEMI-VOWELS:

Liquids
- /r/ ran
- /l/ liquid

Glides
- /w/ want
- /y/ yard

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

AFFRICATES:
- /J/ just
- /C/ channel

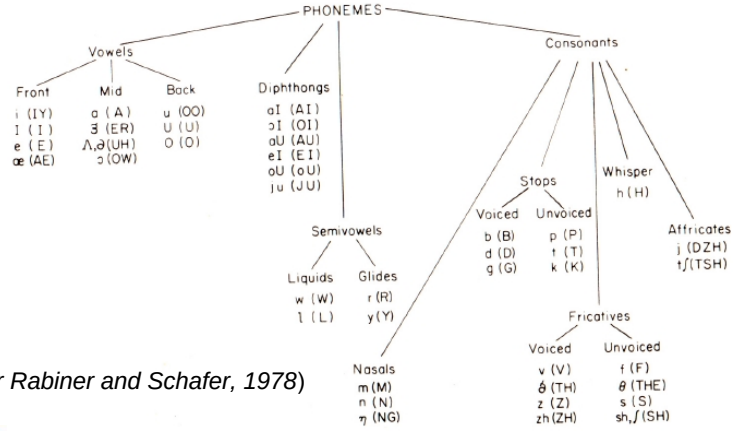
NASALS:
- /m/ mom
- /n/ noon
- /G/ sing

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

(*After J. Hansen*)

Phoneme classification chart

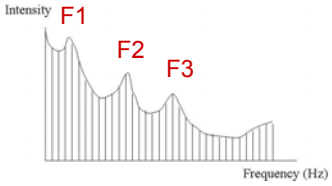
- Sound categorization according to the position of the articulators.



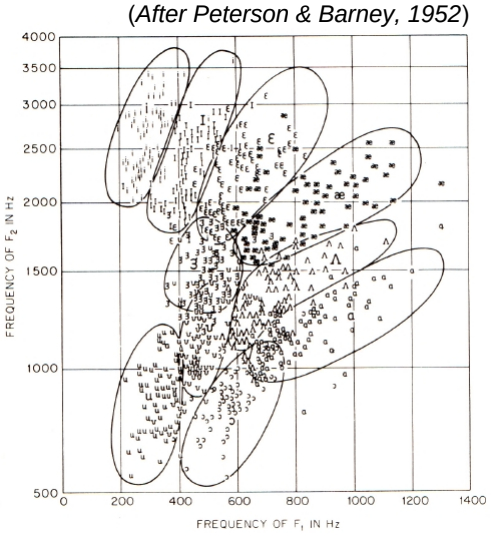
(*After Rabiner and Schafer, 1978*)

The vowel space

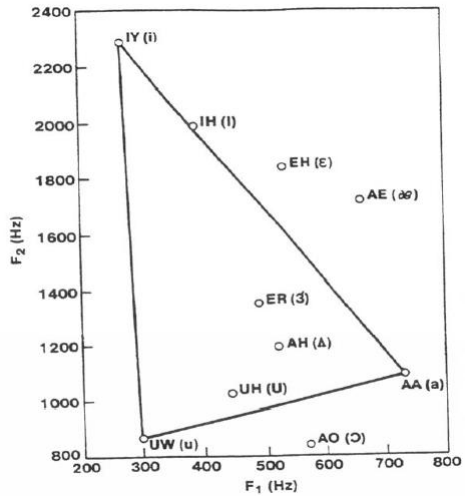
- by the locations of the first and second formant frequencies:



(*After Peterson & Barney, 1952*)



The vowel triangle



Speech Tool

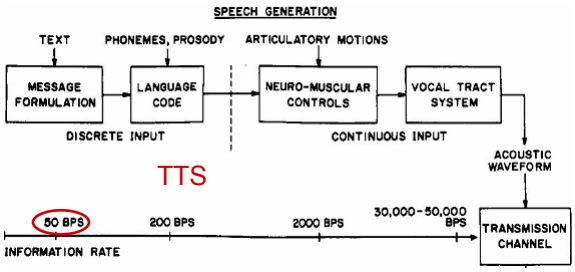
- **Speech Filing System- Tools for Speech Research**
  - It performs standard operations such as recording, replay, waveform editing and labelling, spectrographic and formant analysis and fundamental frequency estimation.
  - http://www.phon.ucl.ac.uk/resource/sfs/

Part III: Speech synthesis

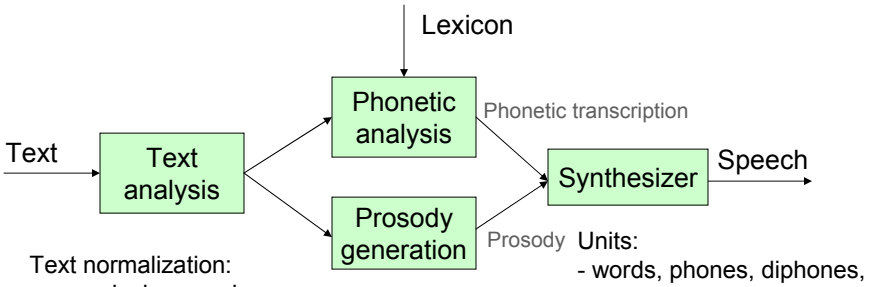
- Introduction
- Basics about speech – a short introduction
- Speech synthesis
  - Articulatory synthesis
  - Formant synthesis
  - Concatenative synthesis

Text-to-speech (TTS)

- TTS converts arbitrary text to intelligible and natural sounding speech.
- TTS is viewed as a speech coding system with an extremely high compression ratio.
- The text file that is input to a speech synthesizer is a form of coded speech. What is the bit rate?



Overview of TTS



Text normalization:
- numerical expansion
- abbreviations
- acronyms
- proper names

Letter-to-sound:
- phonemes
- pitch
- duration
- loudness

Units:
- words, phones, diphones, syllables

Parameters:
- LPC, formants, waveform templates, articulatory

Algorithms:
- rules, concatenation

Text analysis

- Document structure detection
  - to provide context for later processes, e.g. sentence breaking and paragraph segmentation affect prosody.
  - e.g. email needs special care. This is easy :-) ZT
- Text normalization
  - to convert symbols, numbers into an orthographic transcription suitable for phonetic conversion.
  - Dr., 9 am, 10:25, 16/02/2006 (Europe), DK, OPEC
- Linguistic analysis
  - to recover syntactic and semantic features of words, phrases & sentences for both pronunciation and prosodic choices.
  - word type (name or verb), word sense (river or money bank)

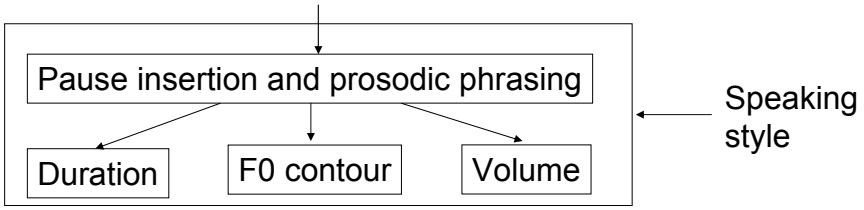
Letter-to-sound

- LTS conversion provides phonetic pronunciation for any sequence of letters.
- Approaches
  - Dictionary lookup
  - If lookup fails, use rules.
    - knight: k -> /sil/ % _n
    - Kitten: k -> /k/
    - Classification and regression trees (CART) is commonly used which includes a set of yes-no questions and a procedure to select the best question at each node to grow the tree from the root.

Prosody

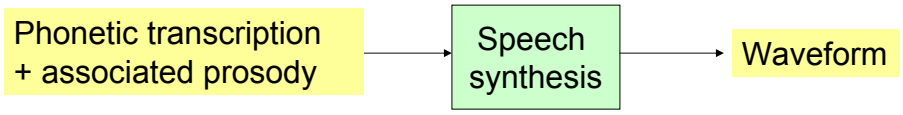
- Pause: indicating phrases and having break
- Pitch: accent, tone, intonation
- Duration
- Loudness

Block diagram of a prosody generation system



Speech synthesis

A module of a TTS system that generates the waveform.



Approaches:

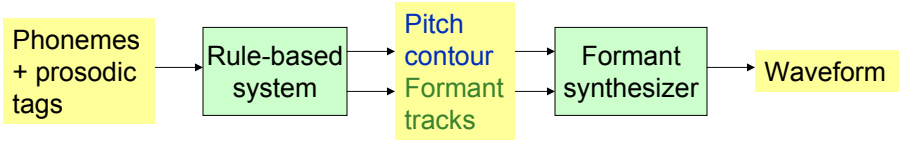
- Limited-domain waveform concatenation, e.g. IVR
- Concatenative systems with no waveform modification, from arbitrary text
- Concatenative systems with waveform modification, for prosody consideration
- Rule-based systems – as opposed to the above data-driven synthesis. For example, formant synthesizer normally uses synthesis by rule.

Types according to the model

- Articulatory synthesis
  - uses a physical model of speech production including all the articulators
- Formant synthesis
  - uses a source-filter model, in which the filter is determined by slowly varying formant frequencies
- Concatenative synthesis
  - concatenates speech segments, where prosody modification plays a key role.

Formant speech synthesis

- A type of synthesis-by-rule where a set of rules are applied to decide how to modify the pitch, formant frequencies, and other parameters from one sound to another
- Block diagram



Concatenative speech synthesis

- Synthesis-by-rule generates unnatural speech
- Concatenative synthesis
  - A speech segment is generated by playing back waveform with matching phoneme string.
    - cut and paste, no rules required
    - completely natural segments
  - An utterance is synthesized by concatenating several speech segments. Discontinuities exist:
    - spectral discontinuities due to formant mismatch at the concatenation point
    - prosodic discontinuities due to pitch mismatch at the concatenation point

Key issues in concatenative synthesis

- Choice of unit
  - Speech segment: phoneme, diphone, word, sentence?
- Design of the set of speech segments
  - Set of speech segments: which and how many?
- Choice of speech segments
  - How to select the best string of speech segments from a given library of segments, given a phonetic string and its prosody?
- Modification of the prosody of a speech segment
  - To best match the desired output prosody

Choice of unit

- Unit types in English (After Huang et al., 2001)

Unit length	Unit type	# units	Quality
Short ↓ Long	Phoneme	42	Low ↓ High
	Diphone	~1500	
	Triphone	~30K	
	Semisyllable	~2000	
	Syllable	~15K	
	Word	100K-1.5M	
	Phrase	∞	
	Sentence	∞	

Attributes of speech synthesis system

- Delay
  - For interactive applications, < 200ms
- Momory resources
  - Rule-based, < 200 KB; Concatenative systems, 100 MB
- CPU resources
  - For concatenative systems, searching may be a problem
- Variable speed
  - e.g., fast speech; difficult for concatenative system
- Pitch control
  - e.g., a specific pitch requirement; difficult for concatenative
- Voice characteristics
  - e.g., specific voices like robot; difficult for concatenative

TTS Systems

- ATT
- Festival

Summary

- Introduction
  - Speech input and output – components of speech interaction
  - State-of-the-art
- Basics about speech – a short introduction
- Speech synthesis
  - Articulatory synthesis
  - Formant synthesis
  - Concatenative synthesis