

## Digital Speech Processing— Lecture 1

### Introduction to Digital Speech Processing

1

## Speech Processing

- Speech is the most natural form of human-human communications.
- Speech is related to language; linguistics is a branch of social science.
- Speech is related to human physiological capability; physiology is a branch of medical science.
- Speech is also related to sound and acoustics, a branch of physical science.
- Therefore, speech is one of the most intriguing signals that humans work with every day.
- Purpose of speech processing:
  - To understand speech as a means of communication;
  - To represent speech for transmission and reproduction;
  - To analyze speech for automatic recognition and extraction of information
  - To discover some physiological characteristics of the talker.

2

## Why Digital Processing of Speech?

- digital processing of speech signals (DPSS) enjoys an *extensive theoretical and experimental base* developed over the past 75 years
- much research has been done since 1965 on the use of *digital signal processing* in speech communication problems
- highly advanced *implementation technology* (VLSI) exists that is well matched to the computational demands of DPSS
- there are *abundant applications* that are in widespread use commercially

3

## The Speech Stack

**Speech Applications** — coding, synthesis, recognition, understanding, verification, language translation, speed-up/slow-down

**Speech Algorithms** — speech-silence (background), voiced-unvoiced decision, pitch detection, formant estimation

**Speech Representations** — temporal, spectral, homomorphic, LPC

**Fundamentals** — acoustics, linguistics, pragmatics, speech perception

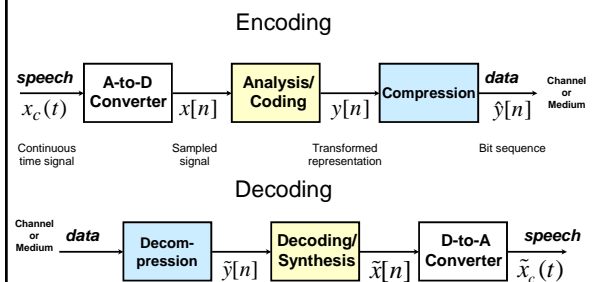
4

## Speech Applications

- We look first at the top of the speech processing stack—namely applications
  - speech coding
  - speech synthesis
  - speech recognition and understanding
  - other speech applications

5

## Speech Coding



6

## Speech Coding

- **Speech Coding** is the process of transforming a speech signal into a representation for efficient transmission and storage of speech
  - narrowband and broadband wired telephony
  - cellular communications
  - Voice over IP (VoIP) to utilize the Internet as a real-time communications medium
  - secure voice for privacy and encryption for national security applications
  - extremely narrowband communications channels, e.g., battlefield applications using HF radio
  - storage of speech for telephone answering machines, IVR systems, prerecorded messages

7

## Demo of Speech Coding

- **Narrowband Speech Coding:**
  - ❖ 64 kbps PCM
  - ❖ 32 kbps ADPCM
  - ❖ 16 kbps LDCELP
  - ❖ 8 kbps CELP
  - ❖ 4.8 kbps FS1016
  - ❖ 2.4 kbps LPC10E
- **Wideband Speech Coding:**
  - Male talker / Female Talker
    - ❖ 3.2 kHz – uncoded
    - ❖ 7 kHz – uncoded
    - ❖ 7 kHz – 64 kbps
    - ❖ 7 kHz – 32 kbps
    - ❖ 7 kHz – 16 kbps



Narrowband Speech



Wideband Speech

8

## Demo of Audio Coding

- CD Original (1.4 Mbps) versus MP3-coded at 128 kbps
  - > female vocal
  - > trumpet selection
  - > orchestra
  - > baroque
  - > guitar

Can you determine which is the uncoded and which is the coded audio for each selection?



Audio Coding



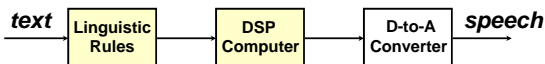
Additional Audio Selections

9

## Audio Coding

- **Female vocal** – MP3-128 kbps coded, CD original
- **Trumpet selection** – CD original, MP3-128 kbps coded
- **Orchestral selection** – MP3-128 kbps coded
- **Baroque** – CD original, MP3-128 kbps coded
- **Guitar** – MP3-128 kbps coded, CD original<sub>0</sub>

## Speech Synthesis




11


## Speech Synthesis

- **Synthesis of Speech** is the process of generating a speech signal using computational means for effective human-machine interactions
  - machine reading of text or email messages
  - telematics feedback in automobiles
  - talking agents for automatic transactions
  - automatic agent in customer care call center
  - handheld devices such as foreign language phrasebooks, dictionaries, crossword puzzle helpers
  - announcement machines that provide information such as stock quotes, airlines schedules, weather reports, etc.

12

## Speech Synthesis Examples

- Soliloquy from Hamlet: 

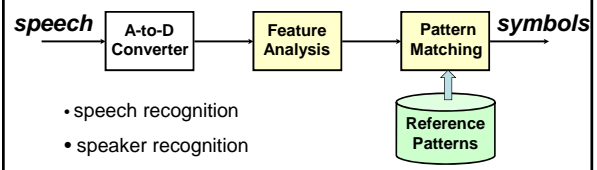
- Gettysburg Address: 

- Third Grade Story:  

1964-lrr      2002-tts

13

## Pattern Matching Problems



- speech recognition
- speaker recognition
- speaker verification
- word spotting
- automatic indexing of speech recordings

14

## Speech Recognition and Understanding

- **Recognition and Understanding of Speech** is the process of extracting usable linguistic information from a speech signal in support of human-machine communication by voice
  - command and control (C&C) applications, e.g., simple commands for spreadsheets, presentation graphics, appliances
  - voice dictation to create letters, memos, and other documents
  - natural language voice dialogues with machines to enable Help desks, Call Centers
  - voice dialing for cellphones and from PDA's and other small devices
  - agent services such as calendar entry and update, address list modification and entry, etc.

15

## Speech Recognition Demos



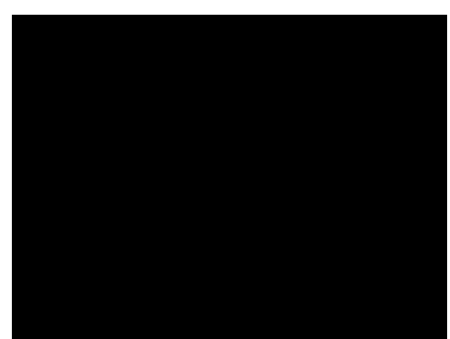
16

## Speech Recognition Demos



17

## Dictation Demo



18

## Other Speech Applications

- **Speaker Verification** for secure access to premises, information, virtual spaces
- **Speaker Recognition** for legal and forensic purposes—national security; also for personalized services
- **Speech Enhancement** for use in noisy environments, to eliminate echo, to align voices with video segments, to change voice qualities, to speed-up or slow-down prerecorded speech (e.g., talking books, rapid review of material, careful scrutinizing of spoken material, etc) => potentially to improve intelligibility and naturalness of speech
- **Language Translation** to convert spoken words in one language to another to facilitate natural language dialogues between people speaking different languages, i.e., tourists, business people

19

## DSP/Speech Enabled Devices



Internet Audio



Digital Cameras



PDA's & Streaming Audio/Video



Cell Phones



Hearing Aids

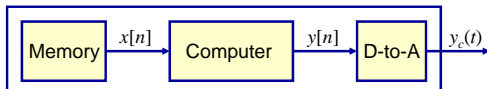


20

## Apple iPod

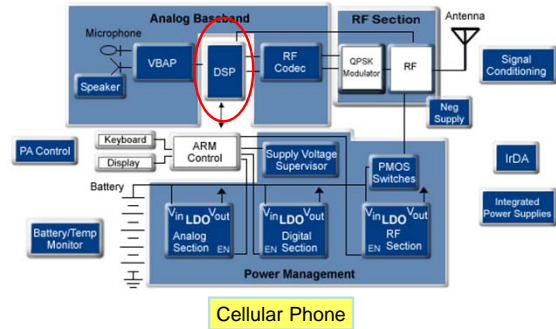


- stores music in MP3, AAC, MP4, wma, wav, ... audio formats
- compression of 11-to-1 for 128 kbps MP3
- can store order of 20,000 songs with 30 GB disk
- can use flash memory to eliminate all moving memory access
- can load songs from iTunes store – more than 1.5 billion downloads
- tens of millions sold



21

## One of the Top DSP Applications



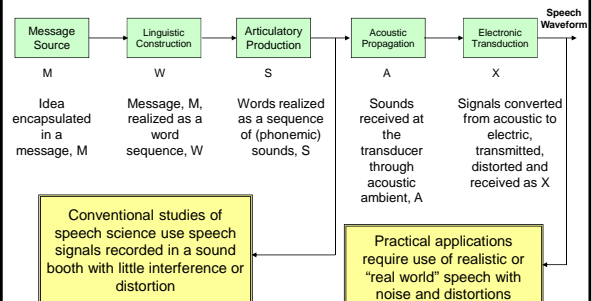
22

## Digital Speech Processing

- Need to understand the **nature of the speech signal**, and how dsp techniques, communication technologies, and information theory methods can be applied to help solve the various application scenarios described above
  - most of the course will concern itself with **speech signal processing** — i.e., converting one type of speech signal representation to another so as to uncover various mathematical or practical properties of the speech signal and do appropriate processing to aid in solving both fundamental and deep problems of interest

23

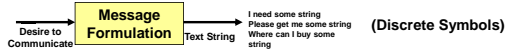
## Speech Signal Production



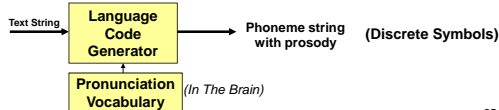
24

## Speech Production/Generation Model

- **Message Formulation** → desire to communicate an idea, a wish, a request, ... => express the message as a sequence of words



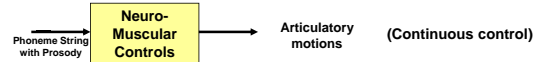
- **Language Code** → need to convert chosen text string to a sequence of sounds in the language that can be understood by others; need to give some form of emphasis, prosody (tune, melody) to the spoken sounds so as to impart non-speech information such as sense of urgency, importance, psychological state of talker, environmental factors (noise, echo)



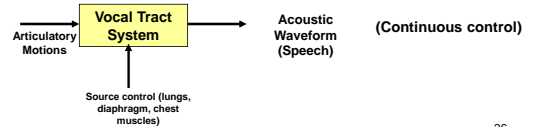
25

## Speech Production/Generation Model

- **Neuro-Muscular Controls** → need to direct the neuro-muscular system to move the articulators (tongue, lips, teeth, jaws, velum) so as to produce the desired spoken message in the desired manner

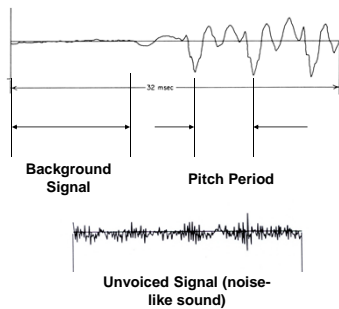


- **Vocal Tract System** → need to shape the human vocal tract system and provide the appropriate sound sources to create an acoustic waveform (speech) that is understandable in the environment in which it is spoken



26

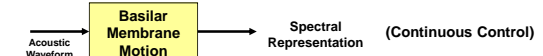
## The Speech Signal



27

## Speech Perception Model

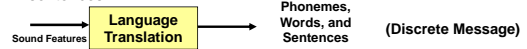
- The acoustic waveform impinges on the ear (the basilar membrane) and is spectrally analyzed by an equivalent filter bank of the ear



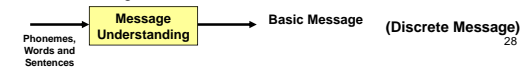
- The signal from the basilar membrane is neurally transduced and coded into features that can be decoded by the brain



- The brain decodes the feature stream into sounds, words and sentences

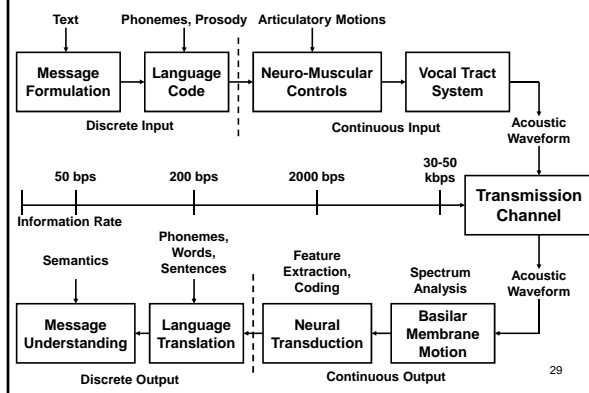


- The brain determines the meaning of the words via a message understanding mechanism



28

## The Speech Chain



29

## The Speech Chain

### Speech Generation Example

Goal: Find out if your office mate has had lunch.

Text: "Did you eat yet?"

Phonemes: "dɪd ju ɪt jɛt?"

Articulator Dynamics: dɪ jə ɪt jɛt

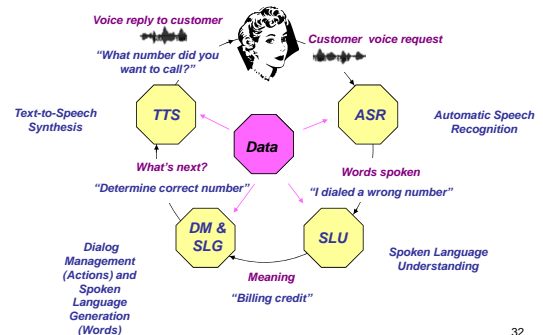
30

## Speech Sciences

- **Linguistics:** science of language, including phonetics, phonology, morphology, and syntax
- **Phonemes:** smallest set of units considered to be the basic set of distinctive sounds of a languages (20-60 units for most languages)
- **Phonemics:** study of phonemes and phonemic systems
- **Phonetics:** study of speech sounds and their production, transmission, and reception, and their analysis, classification, and transcription
- **Phonology:** phonetics and phonemics together
- **Syntax:** meaning of an utterance

31

## The Speech Circle

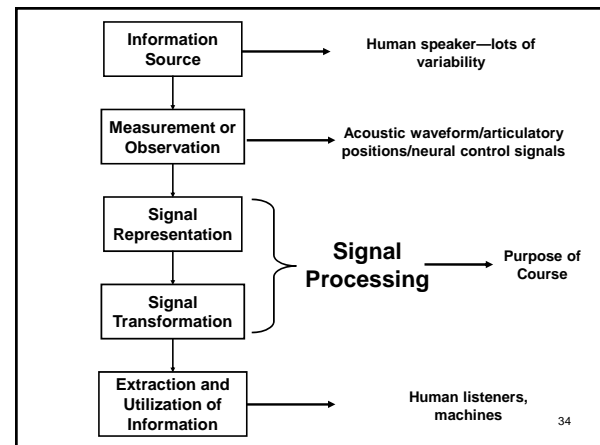


32

## Information Rate of Speech

- from a Shannon view of information:
  - message content/information-- $2^{16}$  symbols (phonemes) in the language; 10 symbols/sec for normal speaking rate => 60 bps is the equivalent information rate for speech (issues of phoneme probabilities, phoneme correlations)
- from a communications point of view:
  - speech bandwidth is between 4 (telephone quality) and 8 kHz (wideband hi-fi speech)—need to sample speech at between 8 and 16 kHz, and need about 8 (log encoded) bits per sample for high quality encoding =>  $8000 \times 8 = 64000$  bps (telephone) to  $16000 \times 8 = 128000$  bps (wideband)

1000-2000 times change in information rate from discrete message symbols to waveform encoding => can we achieve this three orders of magnitude reduction in information rate on real speech waveforms?



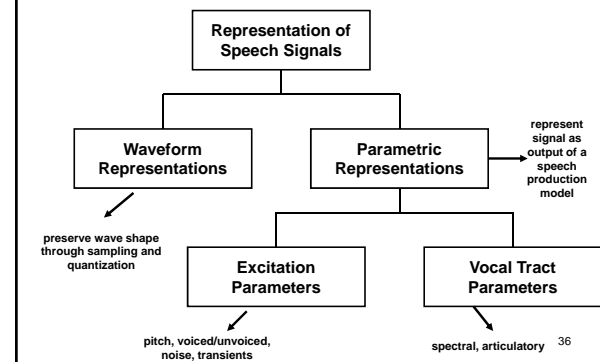
34

## Digital Speech Processing

- DSP:
  - obtaining discrete representations of speech signal
  - theory, design and implementation of numerical procedures (algorithms) for processing the discrete representation in order to achieve a goal (recognizing the signal, modifying the time scale of the signal, removing background noise from the signal, etc.)
- Why DSP
  - reliability
  - flexibility
  - accuracy
  - real-time implementations on inexpensive dsp chips
  - ability to integrate with multimedia and data
  - encryptability/security of the data and the data representations via suitable techniques

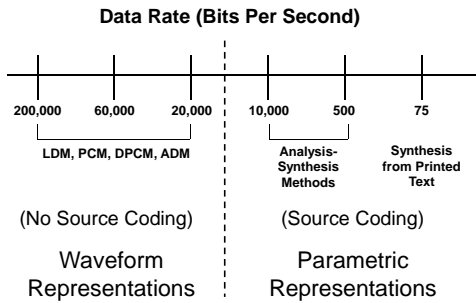
35

## Hierarchy of Digital Speech Processing



36

## Information Rate of Speech



37

## Speech Processing Applications

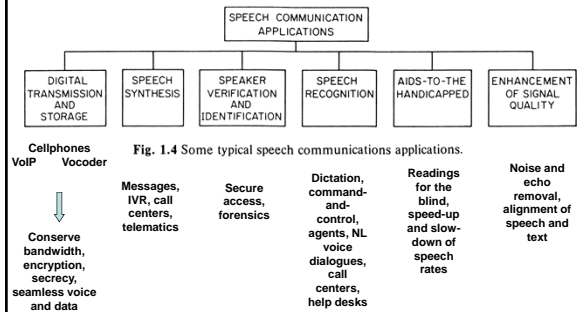
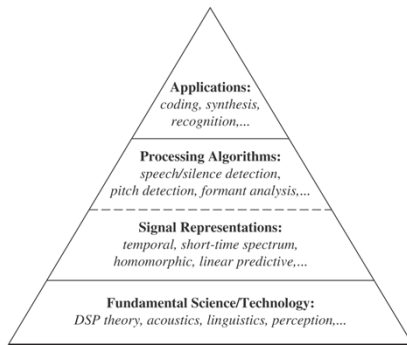


Fig. 1.4 Some typical speech communications applications.

38

## The Speech Stack



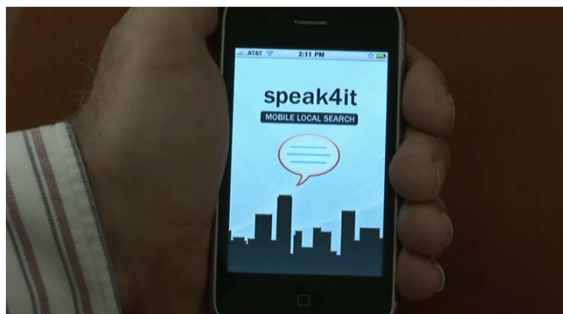
## Intelligent Robot?

<http://www.youtube.com/watch?v=uvQCJpZJH8>



40

## Speak 4 It (AT&T Labs)



Courtesy: Mazin Rahim

41

## What We Will Be Learning

- review some basic dsp concepts
- speech production model—acoustics, articulatory concepts, speech production models
- speech perception model—ear models, auditory signal processing, equivalent acoustic processing models
- time domain processing concepts—speech properties, pitch, voiced-unvoiced, energy, autocorrelation, zero-crossing rates
- short time Fourier analysis methods—digital filter banks, spectrograms, analysis-synthesis systems, vocoders
- homomorphic speech processing—cepstrum, pitch detection, formant estimation, homomorphic vocoder
- linear predictive coding methods—autocorrelation method, covariance method, lattice methods, relation to vocal tract models
- speech waveform coding and source models—delta modulation, PCM, mu-law, ADPCM, vector quantization, multipulse coding, CELP coding
- methods for speech synthesis and text-to-speech systems—physical models, formant models, articulatory models, concatenative models
- methods for speech recognition—the Hidden Markov Model (HMM)

42