# Lecture 01:
# [Rabiner] Chapter 1
# Introduction to Digital Speech Processing

DEEE725 음성신호처리실습
Speech Signal Processing Lab
Instructor: 장길진

# Introduction to Speech Processing

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

# Basics

- ***speech*** is composed of a sequence of sounds
- ***sounds*** (and transitions between them) serve as a symbolic representation of information to be shared between humans (or humans and machines)
- arrangement of sounds is governed by rules of ***language*** (constraints on sound sequences, word sequences, etc.) – /spl/ exists, /sbk/ doesn’t exist
- ***linguistics*** is the study of the rules of language
- ***phonetics*** is the study of the sounds of speech

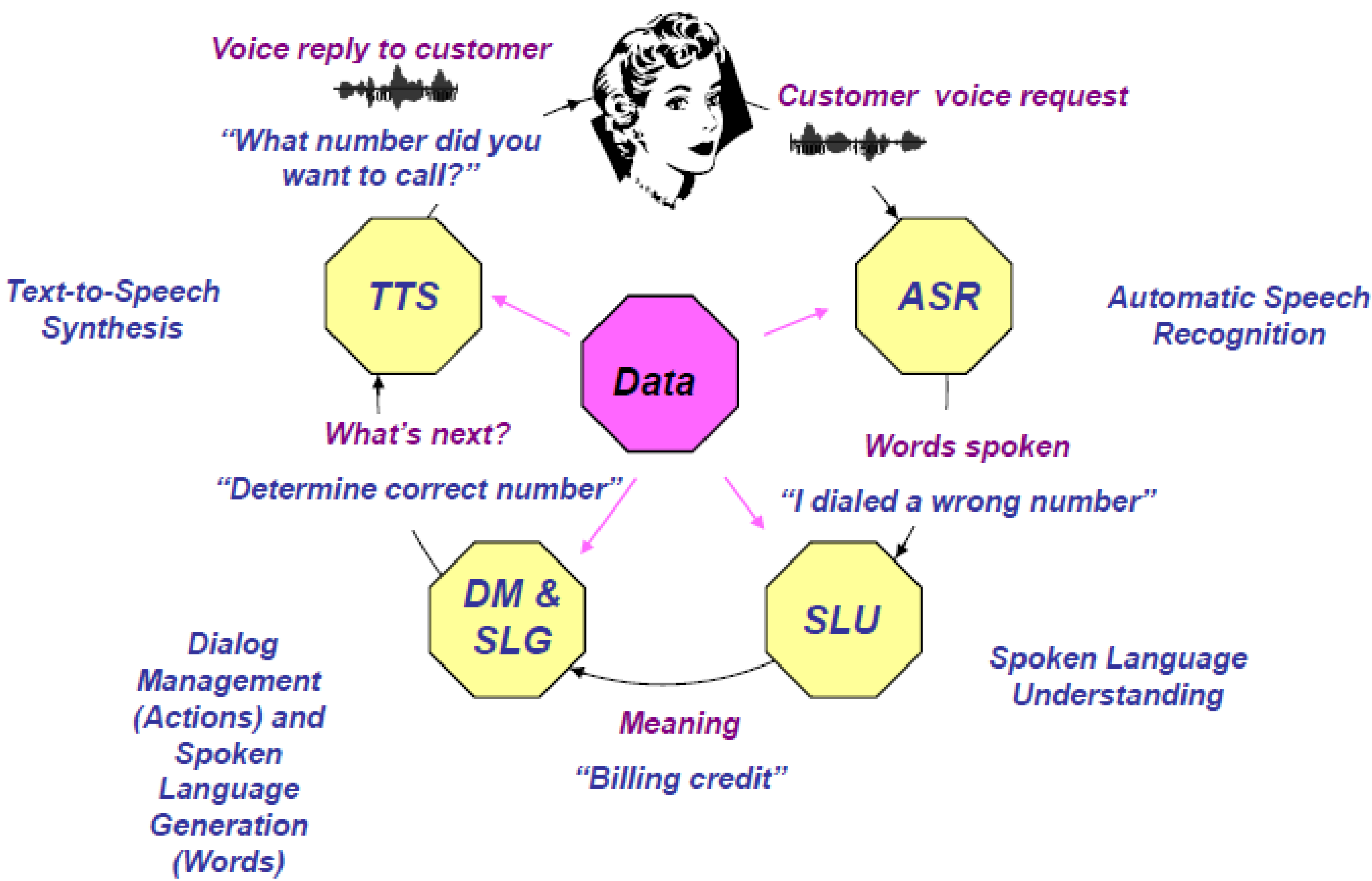
# Speech Sciences

- 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

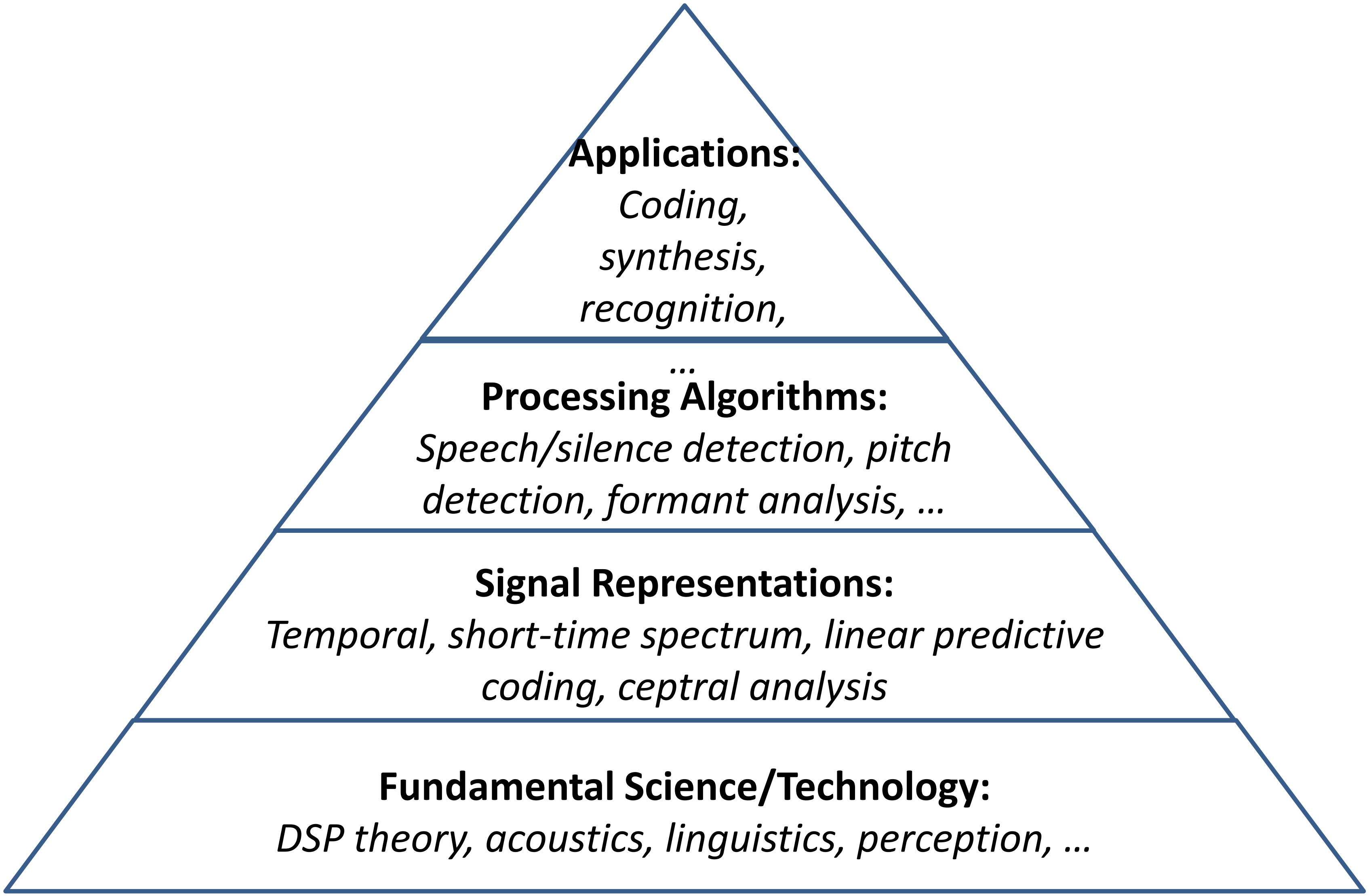
# Speech Applications

- Prevalent speech applications are
  - speech coding (vocoder)
  - speech synthesis (Text-to-speech; TTS)
  - speech recognition and understanding (Speech-to-text; STT)
  - other applications

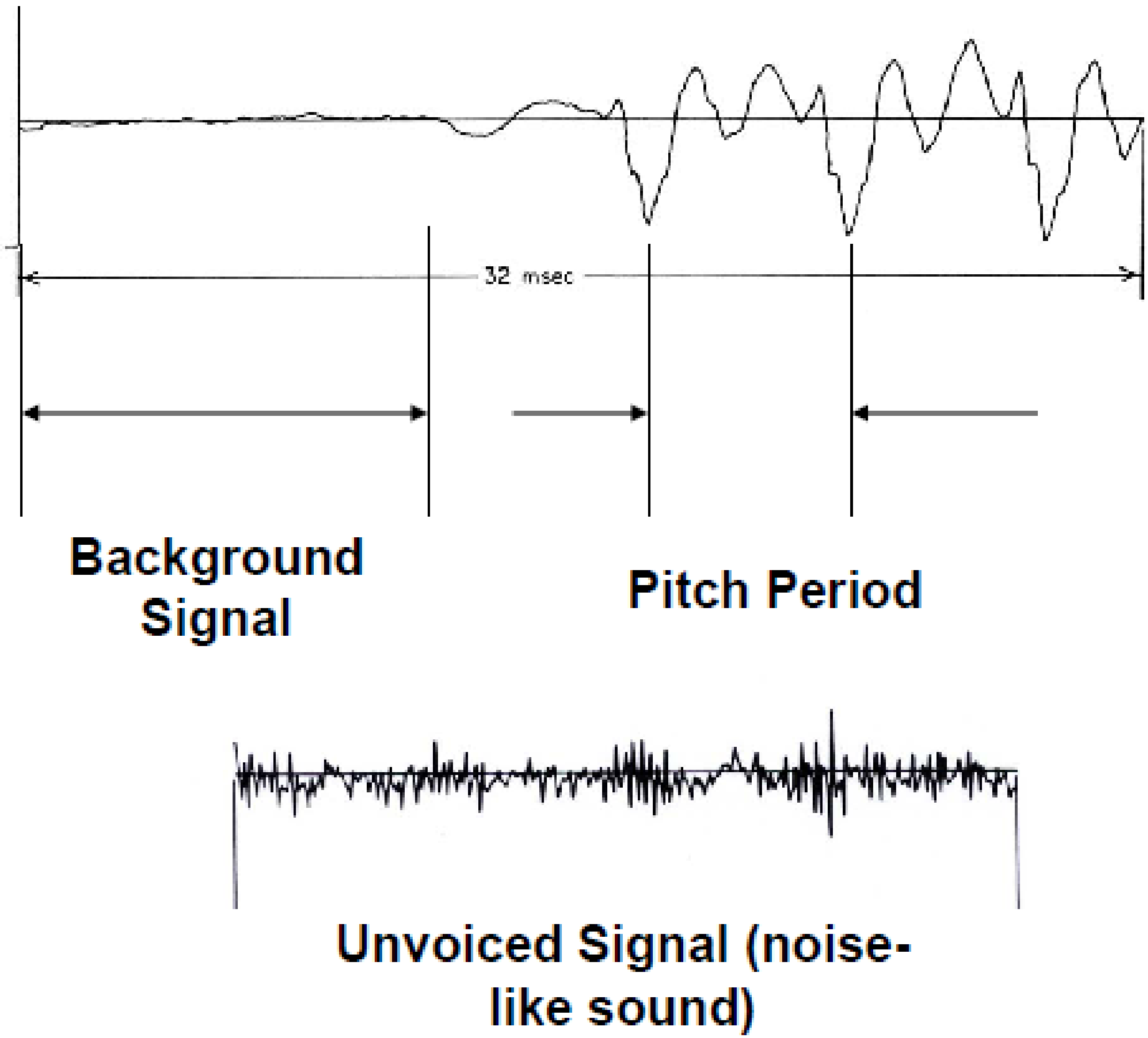
# The Speech Cycle



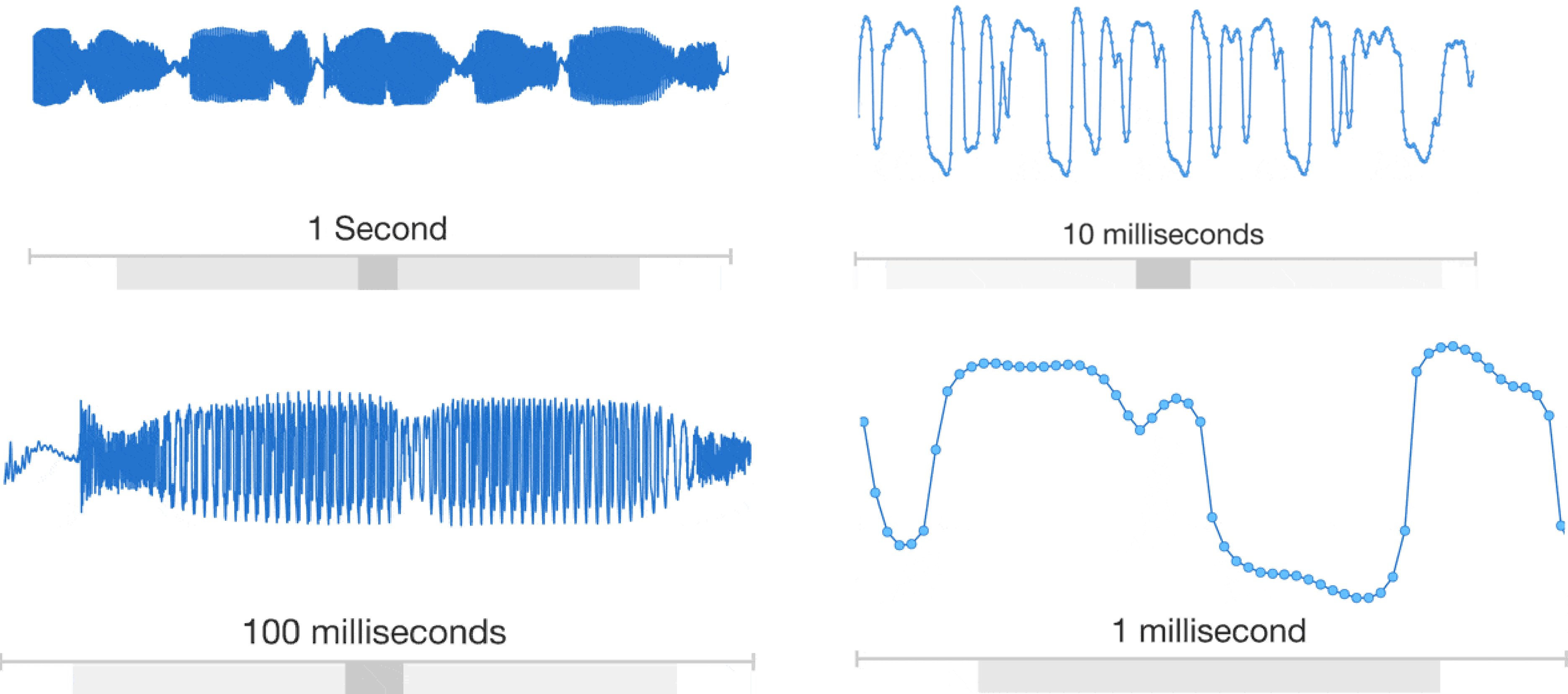
# The Speech Stack



# The Speech Signal



# Speech Signals in Various Scales




*Slide credit: Yuchen Fan, Matt Potok, Christopher Shroba*

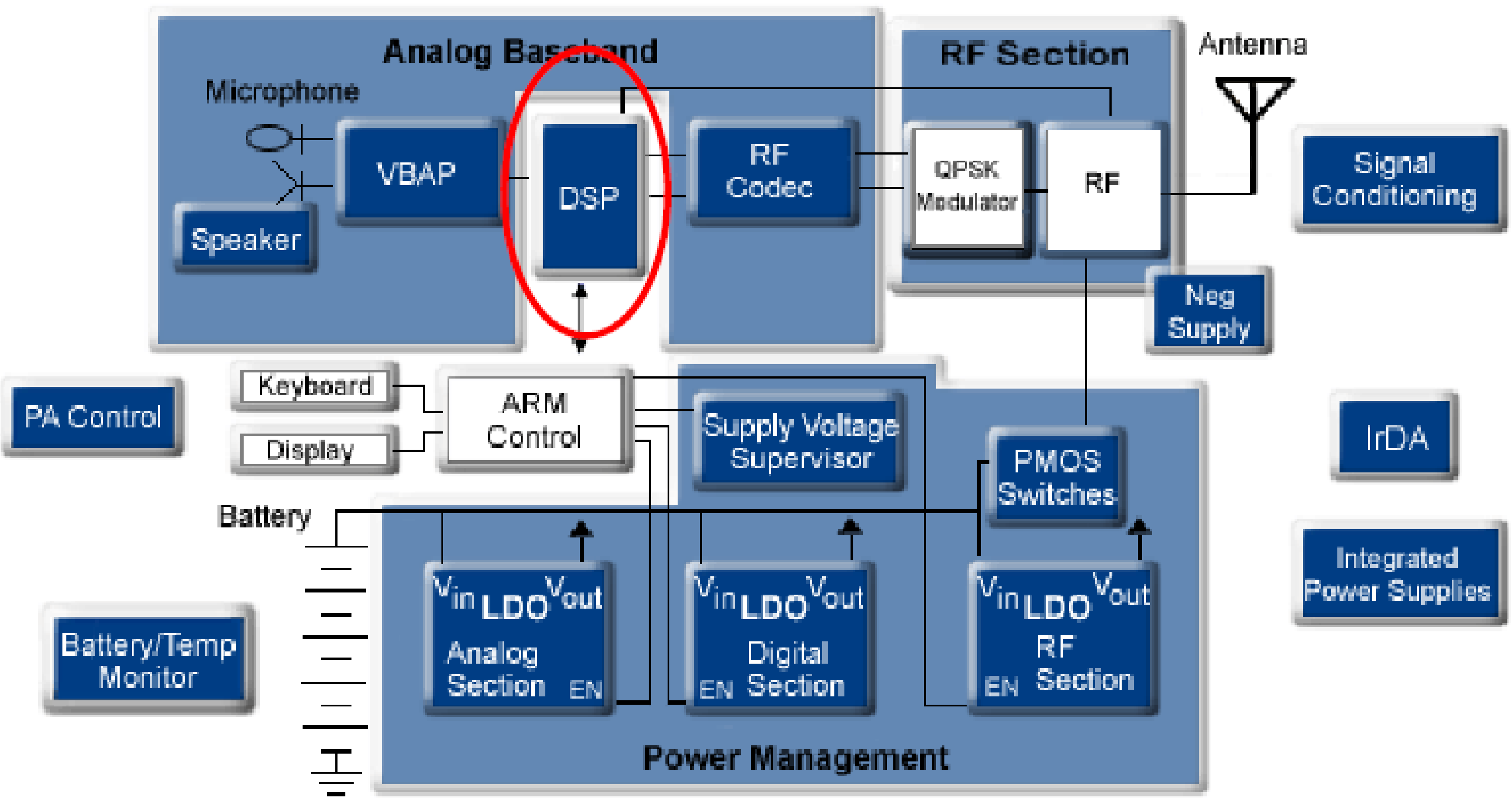
# Digital Speech Processing

- 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

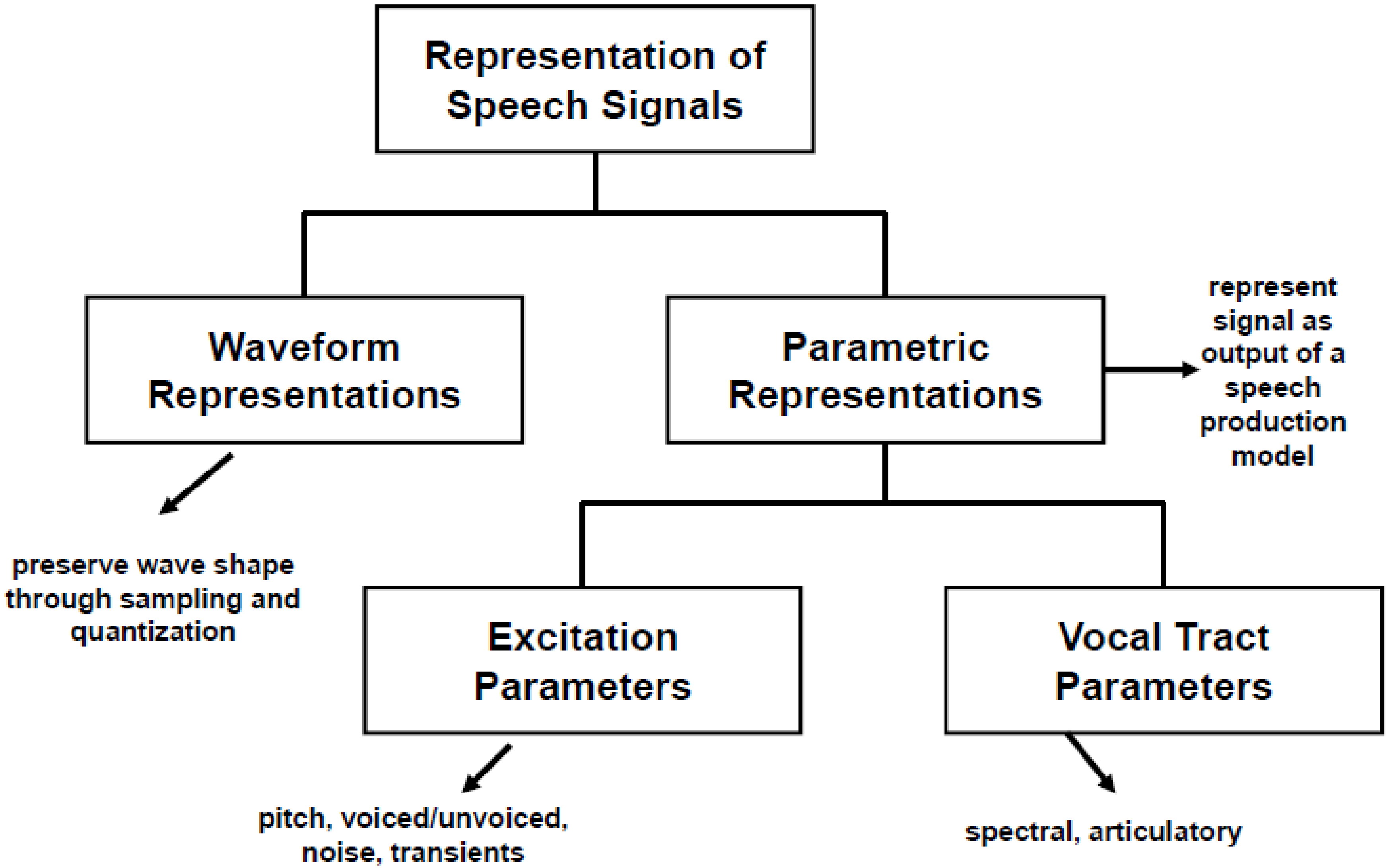
# Why Digital Processing of Speech?

- ***digital processing of speech signals (DPSS)*** enjoys an ***extensive theoretical and experimental base*** developed over the past 80 years
- much research has been done since 1965 on the use of ***digital signal processing* (*DSP*)** 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 and commercial uses

# A Cellular Phone –
# One of the Top DSP Applications

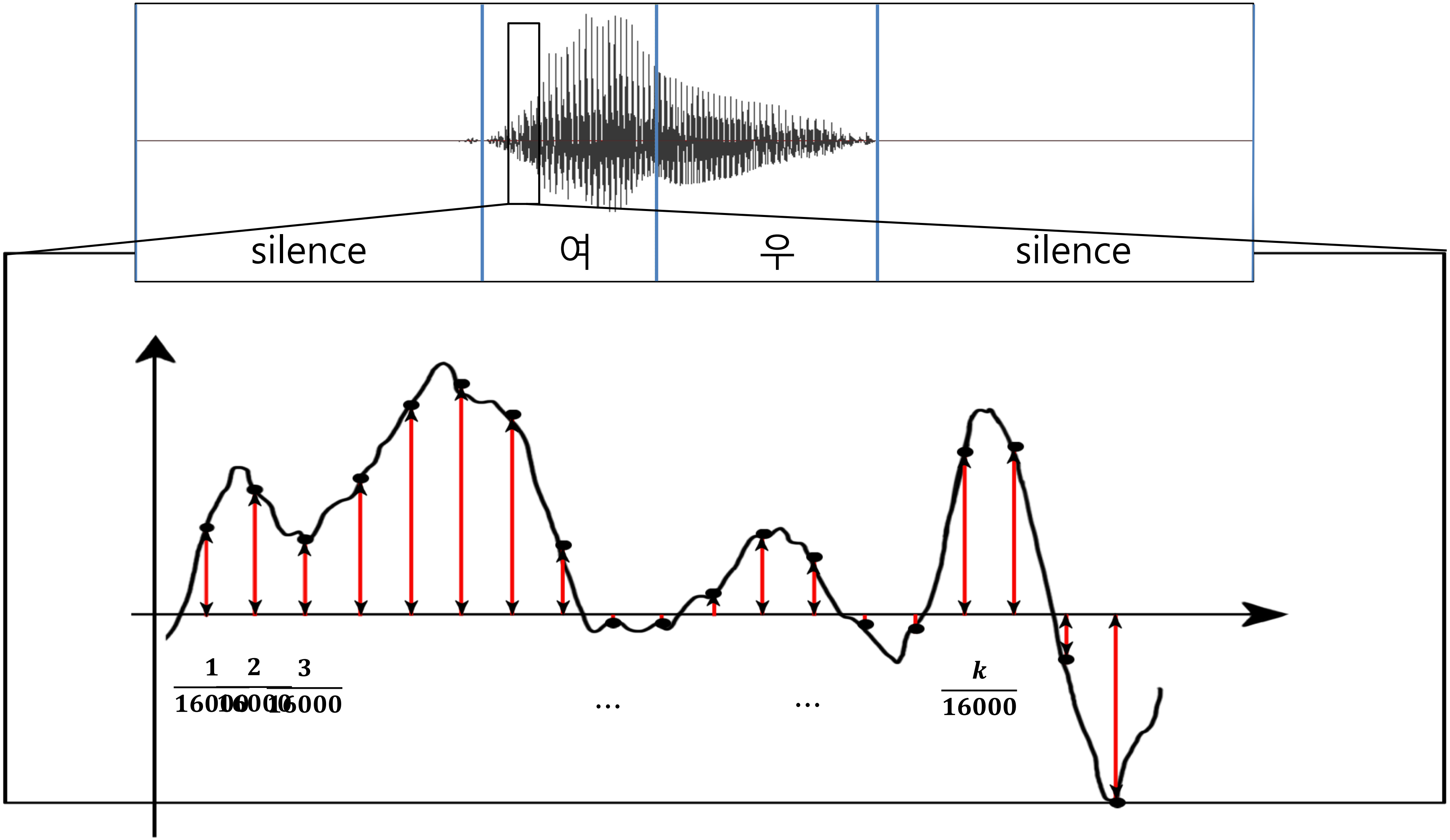


# Hierarchy of Digital Speech Processing



# Digital Speech Representation

Sampling rate: 16K, PCM (pulse code modulation, Analog-to-digital)






*Slide credit: Jihwan Kim Sogan University*

# PCM (pulse code modulation)

- Sampling rate (Fs, Hz)
  - Number of samples per unit time (usually second)
    - According to Nyquist theorem, up to Fs/2 Hz can be represented in the frequency domain
  - Example
    - Music CD: 44100 Hz = **44.1** kHz
      - Human-audible frequency range is known to be 20Hz~20kHz
    - Vocoders including 2G cellular phone: **8** kHz
    - Speech recognition: **8** kHz → **16** kHz
  - Number of bytes per sample: **2** bytes = **16** bits
    - $2^{16} = 65{,}536$ different levels are represented

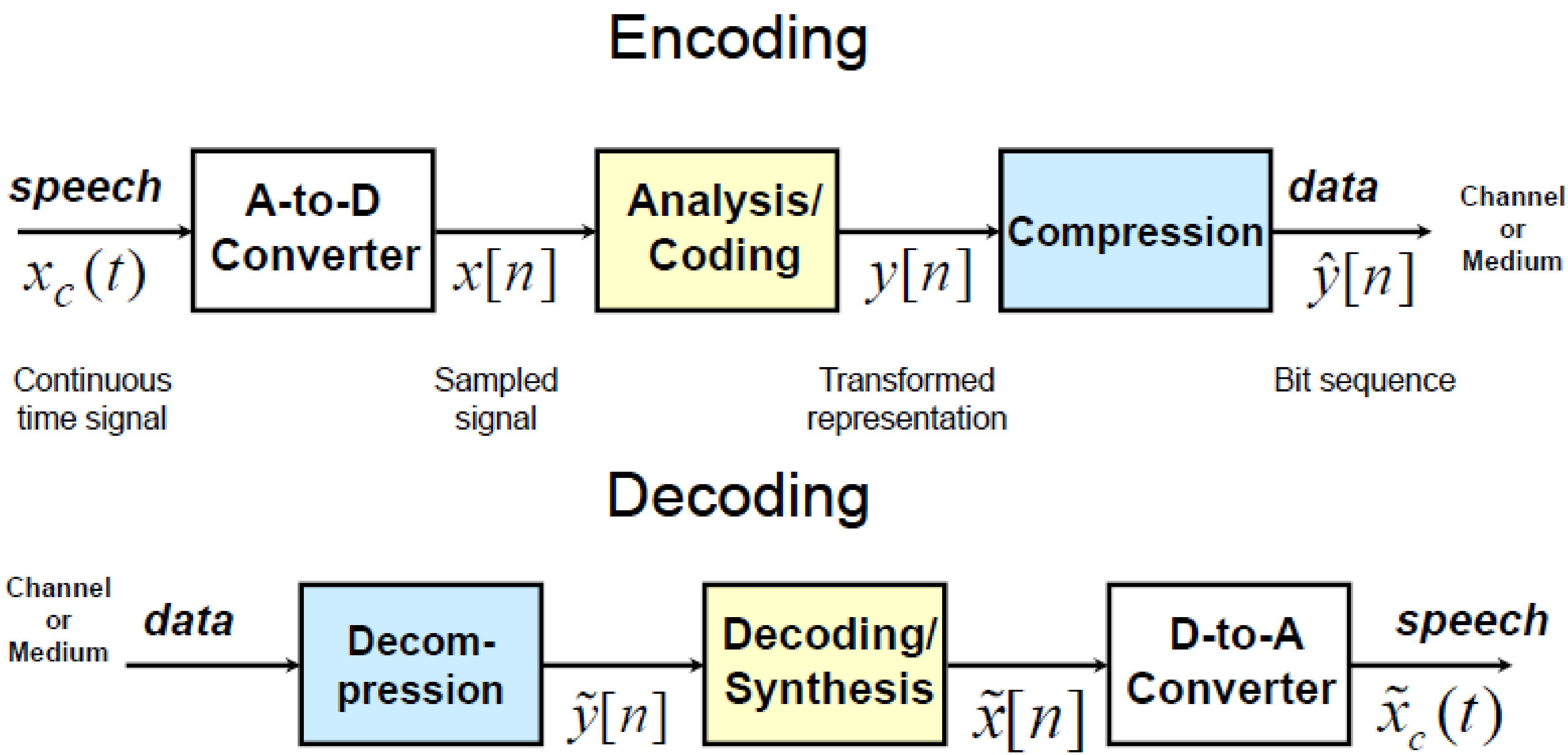

*Slide credit: Jihwan Kim Sogan University*

# Various Data Rates

- CD, 44.1 kHz
  - 1 second = 44.1 * 1000 * 2 bytes = 88,200 bytes = 705,600 bits → 705.6 kbps (kilobits per second)
    - 700MB CD = 700*1^6*8 / 705.6
      = 132 min (mono) or 66 min (stereo)
  - Standard mp3 encoding bitrate is 128 kbps for stereo signal
    - About 1/11 compression
- Vocoders, 8 kHz
  - 1 second = 8 * 1000 * 2 bytes = 16,000 bytes → 128 kbps
- High-quality voice, 16 kHz
  - 1 second = 32,000 bytes → 256 kbps


*Slide credit: Jihwan Kim Sogan University*

# Speech Coding



# Speech Coding

- The process of transforming a speech signal into a representation for ***efficient transmission*** and ***storage***
  - narrowband and broadband wired telephony
  - cellular communications
  - Voice over IP (VoIP) to utilize the Internet as a real-time communications medium
  - extremely narrowband communications channels
    - e.g. battlefield applications using HF radio
- Example coding methods
  - 64 kbps PCM (pulse-code modulation)
  - 32 kbps ADPCM (adaptive differential PCM)
  - 8 kbps CELP (code-excited linear prediction)
  - 2.4 kbps LPC10E
  - less than 1.0 kbps MBE (multi-band excitation)

Slide credits:

Yuchen Fan, Matt Potok, Christopher Shroba

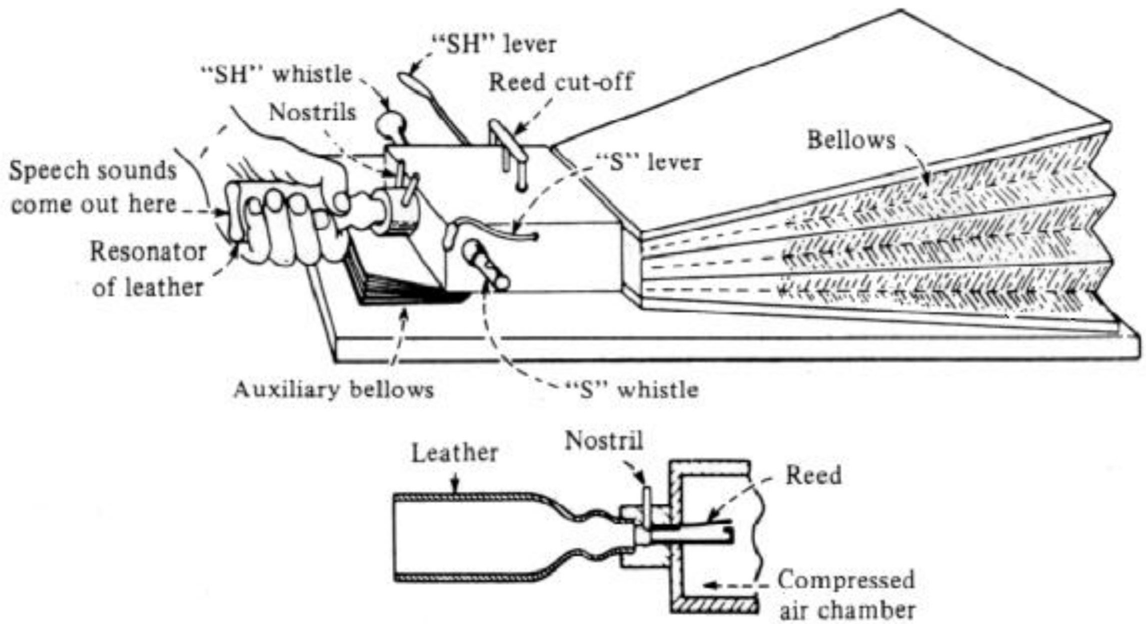
# SPEECH SYNTHESIS: SHORT HISTORY

# Speech Synthesis


- 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
  - handheld devices such as foreign language
- Already, widely used in many applications
  - Try Googling "TTS"

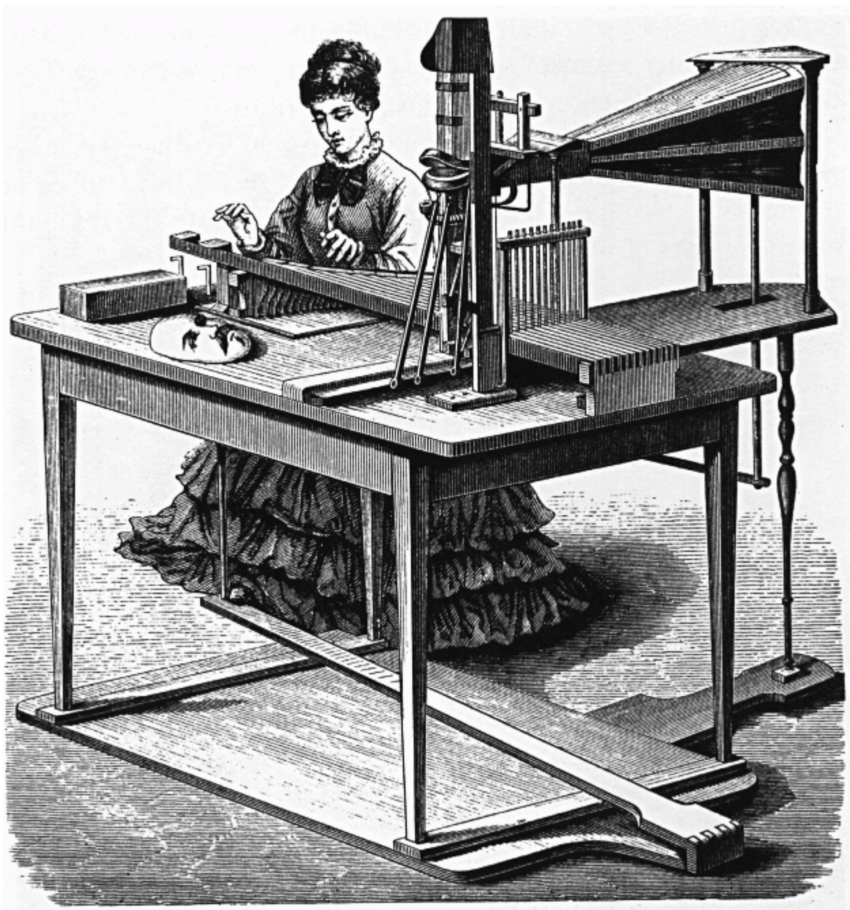
# The First ‘Speaking Machine’

- Wolfgang von Kempelen, Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine, 1791
(in Deutsches Museum still and playable)



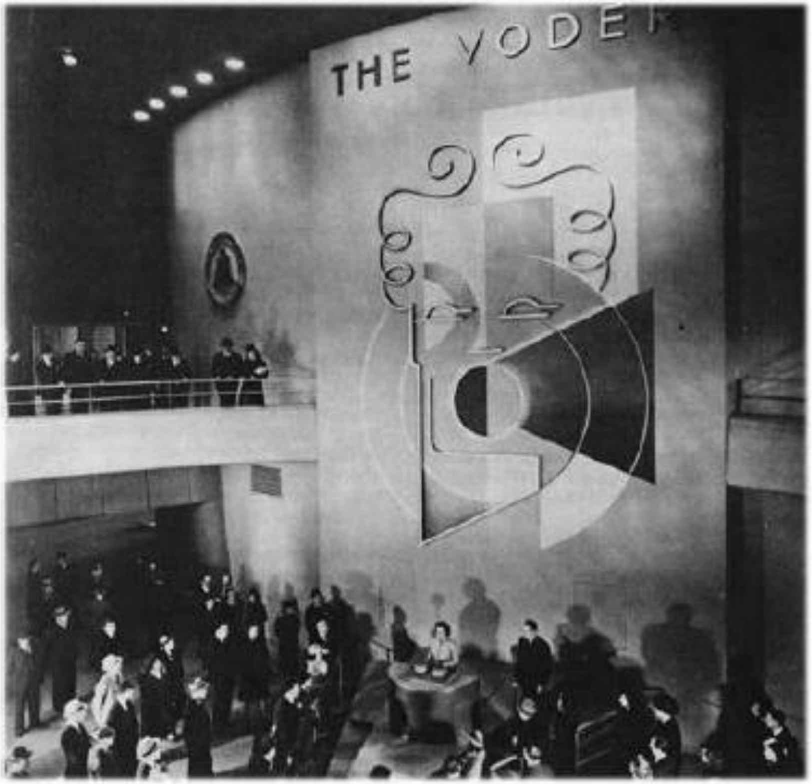

- First to produce whole words, phrases – in many languages

# Joseph Faber's Euphonia, 1846



- Constructed 1835 w/pedal and keyboard control
  - Whispered and ordinary speech
  - Model of tongue, pharyngeal cavity with manipulatable shape
  - Singing too: "God Save the Queen"
- Forerunners of Modern Articulatory Synthesis: George Rosen's DAVO synthesizer (1958) at MIT

# The Voder ...

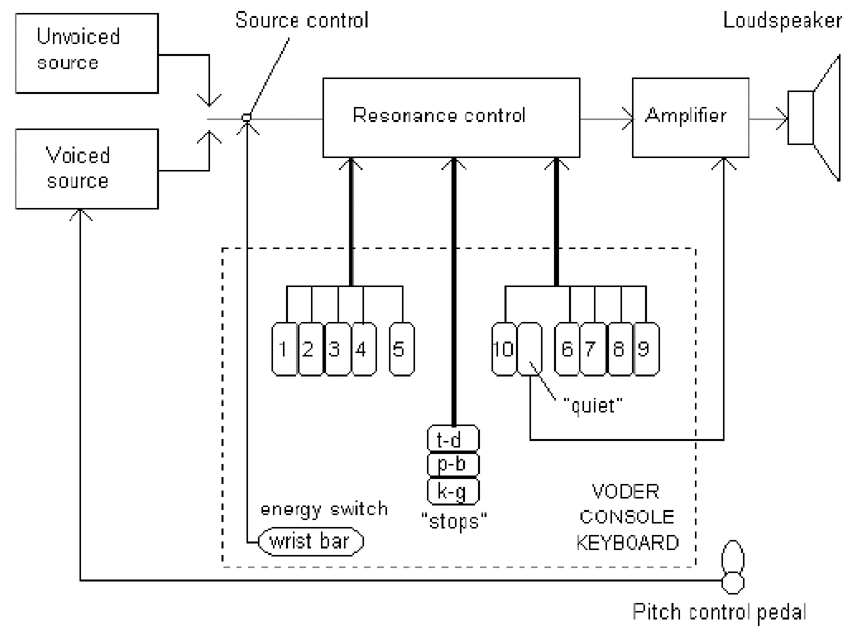



Developed by Homer Dudley at Bell Telephone Laboratories, 1939

# The Voder

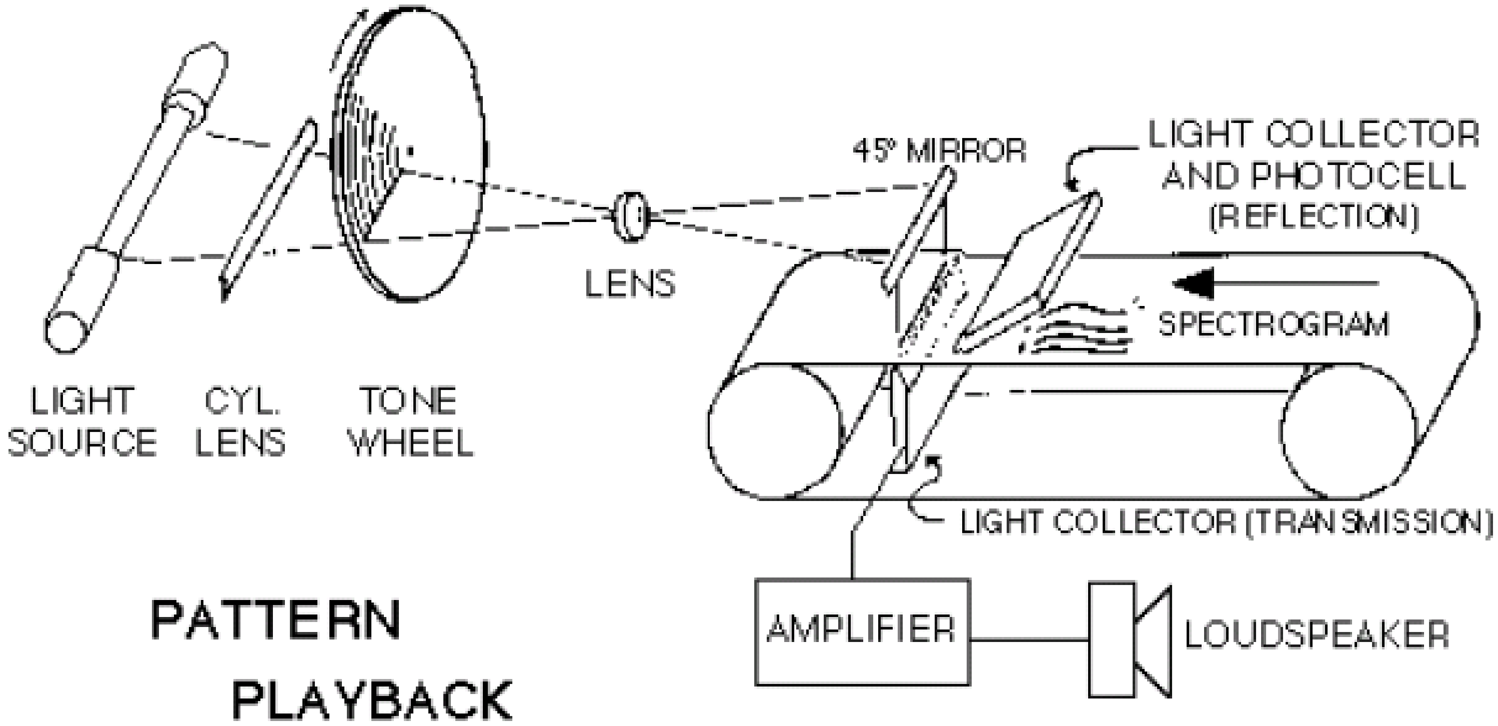
- World's Fair in NY, 1939
- Requires much training to 'play'
- Purpose: coding/compression
  - Reduce bandwidth needed to transmit speech, so many phone calls can be sent over single line

… an acoustic synthesizer




Architectural blueprint for the Voder

# The Pattern Playback




Developed by Franklin Cooper at Haskins Laboratories, 1951

# The Pattern Playback

- Answers:
  - These days a chicken leg is a rare dish.
  - It's easy to tell the depth of a well.
  - Four hours of steady work faced us.
- 'Automatic' synthesis from spectrogram – but can also use hand-painted spectrograms as input
- Purpose: understand perceptual effect of spectral details

# Formant/Resonance/Acoustic Synthesis

- Parametric or resonance synthesis
  - Specify minimal parameters, e.g. f0 and first 3 formants
  - Pass electronic source signal thru filter
    - Harmonic tone for voiced sounds
    - Aperiodic noise for unvoiced
    - Filter simulates the different resonances of the vocal tract
- E.g.
  - Walter Lawrence’s Parametric Artificial Talker (1953) for vowels and consonants
  - Gunnar Fant’s Orator Verbis Electris (1953) for vowels
  - Formant synthesis download (M$demo)

# Concatenative Synthesis

- Most common type today
- First practical application in 1936: British Phone company's Talking Clock
  - Optical storage for words, part-words, phrases
  - Concatenated to tell time
- E.g.
  - And a 'similar' example from Radio Free Vestibule (1994)
  - Bell Labs TTS (1977)    (1985)

# Pronunciation Issues

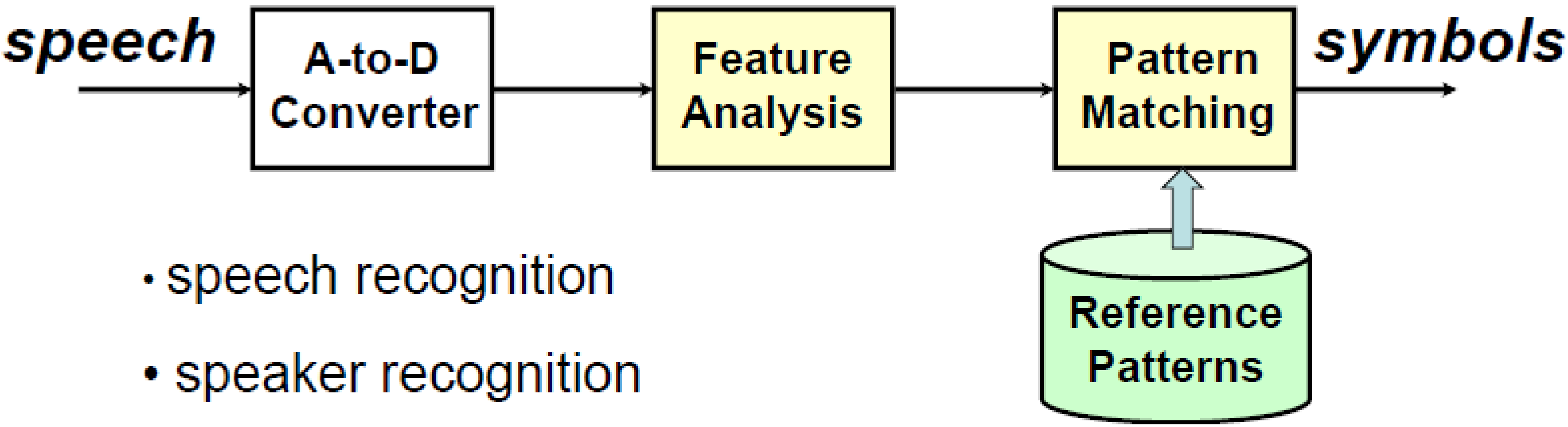
- Rules for disambiguation in context: bass
- Lexicon: comb, tomb, Punxsutawney Phil
  - Letter-to-Sound Rules
    - Hand built
    - Learned from data (pronunciation dictionary)
    - Hard to get good accuracy and coverage – many exceptions
  - Dictionary of pronunciations
    - More accurate
    - New (Out-of-Vocabulary) words a problem

# Not Quite There

- Festival concatenative:
- Acuvoice concatenative:
- HMM synthesis (Rob Donovan):
- Rhetorical unit selection
  - (acquired by Nuance)
- AT&T Labs Naturally Speaking

# SPEECH RECOGNITION

# Pattern Matching Problems




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

# Speech Recognition and Understanding

- 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
  - 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 smartphones
  - voice-driven Internet search
  - **Chatbots**

# Text-to-Phoneme Conversion

- Goal: Find out if your office mate has had lunch already.
- Text: “Did you eat yet”
- Phonemes: /dId yu it yɛt/
- Articulator Dynamics: /dI jə it jɛt/

# Demos (1): Nuance

- https://youtu.be/oNc2f2BhZ50
  - IDF 2012: Nuance Speech Recognition Demo with Nuance Dragon and an Ultrabook at IDF 2012 in San Francisco.
- https://youtu.be/WvbNBBh_wPw
  - Nuance Speech Recognition Demo for EMA
- https://youtu.be/NRE77lW5I2Y
  - How Nuance's Dragon NaturallySpeaking speech recognition software works

# Demos (2)

- https://youtu.be/dXHZqUiManw
  - Jarvis on Ubuntu using Speech Recognition
- https://youtu.be/94IOUW0EQyg
  - pyJARVIS: Ubuntu voice control with python
  - It is a multi-language voice control system developed in python, using natural language processing.
  - GitHub link: https://github.com/rcorcs/NatI
- https://youtu.be/u9FPqkuoEJ8
  - Siraj Raval, “How to Make a Simple Tensorflow Speech Recognizer”
  - Code: https://github.com/llSourcell/tensorflow_speech_recognition_demo

# Demos (3)

- https://youtu.be/NaqZkV_fBIM
  - Neon’s implementation of Baidu’s "Deep Speech 2" model for speech recognition trained on audio-books from the Librispeech corpus.
  - Spectrogram (top left), raw audio (top right), and FFT spectrum (bottom)
- https://youtu.be/g-sndkf7mCs (92 minutes)
  - Deep Learning for Speech Recognition (Adam Coates, Baidu)

# Other Speech Applications

- Speaker Verification
  - secure access to premises, information, and virtual spaces
- Speaker Recognition
  - legal and forensic purposes; also for personalized services
- Speech Enhancement
  - for use in noisy environments, or to eliminate echo
- Voice Conversion
  - 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

# Speech Applications: Summary

| Research field | Tech. level | Relevant tech/theory | Applications |
|---|---|---|---|
| Speech coding | Saturated | Signal processing; Compression; Information theory; Communication | Vocoders; VoIP |
| Speech enhancement / BSS (blind signal separation) | Moderate; Difficult in real conditions | Noisy speech recognition; Echo elimination; Far-field speech recognition; Vocoders | |
| Voice conversion | Moderate | Video/voice alignments; Voice intelligibility and naturalness improvement; talking books, rapid review of material, careful scrutinizing of spoken material, etc. | |
| Speech synthesis | Saturated ➔ Advancing | Natural language processing (NLP); Search | Text-to-speech (TTS) |
| Speech recognition | Difficult; large-scale | Machin learning; Pattern classification; NLP; Deep learning; Artificial intelligence | HCI; ARS; Chatbot; AI speakers |
| Keyword spotting | Moderate; Difficult in real conditions | Machin learning; Pattern classification; ~~NLP;~~ Deep learning; Artificial intelligence | HCI; AI speakers |
| Speaker recognition / verification | Moderate; Difficult in real conditions | Authentication; legal and forensic purposes; personalized services | |
| Translation | Difficult; large-scale | Speech recognition; NLP; Deep learning; RNN | Touring, etc. |

# Topics to be Covered

- Speech production model—acoustics, articulatory concepts, speech production models
- Speech perception model—ear models, auditory signal processing, equivalent acoustic processing models
- Review some basic DSP concepts
- Time domain processing concepts—speech properties, pitch, voiced/unvoiced, energy, autocorrelation, zero-crossing rates
- Short time Fourier analysis methods—digital filter banks, spectrograms, formant estimation
- Linear predictive coding methods—autocorrelation method, covariance method, relation to vocal tract models
- ~~Speech waveform coding and source models—delta modulation, PCM, ADPCM, vector quantization, CELP coding~~
- Speech recognition—the Hidden Markov Model (HMM)
- Deep learning methods for speech recognition (TBA)

DEEE725 음성신호처리실습
장길진

# END OF CHAPTER 1. INTRODUCTION TO DIGITAL SPEECH PROCESSING