EE6424 Digital Audio Signal Processing (Part 2: Speech signal processing)

Dr. Lin Zhiping, Assoc Professor

Office: **S2-B2a-14**

Tel: **6790 6857**

E-mail: **ezplin@ntu.edu.sg**

Topics covered in EE6424 Part 2

- Lecture 1: Introduction to Digital Speech Processing
- Lecture 2: Principal characteristics of speech
- Lecture 3: Sampling and quantization
- Lecture 4: Speech coding
- Lecture 5: Analysis and processing of speech
- Lecture 6: Speech and speaker recognition

Textbooks

- L. R. Rabiner and R. W. Schafer, *Introduction to Digital Speech Processing*. *Foundations and Trends in Signal Processing*, vol. 1, no. 1-2, 2007.

- L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing. Pearson, 2011.

References

- S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. Taylor & Francis, 2001.

- J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-Interscience, 2000.

- T. F. Quatieri, *Discrete-Time Speech Signal Processing – Principles and Practice*. Pearson, 2002.

- D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice-Hall, 2000.

EE6424 Digital Audio Signal Processing
Part 2
Lecture 1:
Introduction to Digital Speech Processing

Outline of lecture

- Digital speech processing
  - The invention of telephone
  - Information rate of speech
  - General speech processing model
- Applications
  - Speech coding
  - Speech recognition
  - Speech-to-speech translation
  - Speaker verification, STARhome@Fusionopolis
  - Spoken language recognition

EE6424 Part 2: Lecture 1.1

DIGITAL PROCESSING OF SPEECH

From analog to digital processing

- "Mr. Watson – Come here – I want to see you" were the first few words Alexander Graham Bell said using his new invention – the telephone, a device that could "telegraph" any sound even the sound of speech. The invention was conceived in 1876.

Bell at the opening of the long-distance line from New York to Chicago in 1892.

IPHONE 5, 2012

- The invention of telephone marked the first step in speech processing. Speech signal is converted from acoustic to electrical waveform, transmitted, and converted back acoustic form by the telephone handset.

- The invention of pulse code modulation (PCM) in 1938 and the subsequent development of digital circuits have made possible digital processing of speech. Since then, the progress in digital speech processing has been remarkable, especially after 1960.

- From speech processing perspective, digital implementation of algorithms is easier and cheaper (usually software instead of hardware, cross-platform, cloud based), and more flexible (re-configurable) compared to an analog system.

Information rate of speech

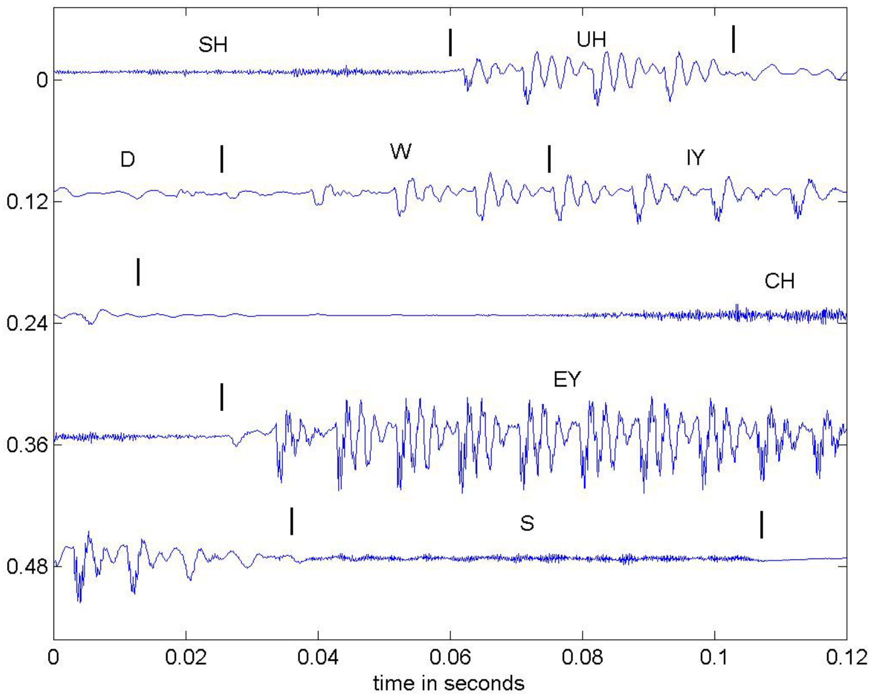
- Shannon's information theory – a message represented as a sequence of discrete symbols can be quantified by its information content in bits, and the rate of transmission of information is measured in bits per second (bps).

- Speech can be represented in terms of its message content, or information. A speech signal is an acoustic waveform carrying the message information.

- The information that is conveyed through speech is intrinsically of a discrete nature. It can be represented by a concatenation of elements from a finite set of symbols, called phonemes. 音素

- Each language has its own distinctive set of phonemes, typically numbering from 30 to 50. For example, English can be represented by a set of around 42 phonemes.

- Figure below shows a speech waveform with phonetic labels for the text message “should we chase”. Notice that the 0.6 second of speech signal carries 8 phones in this specific example.



- ARPAbet is used here, another option is to use the International Phonetic Alphabet (IPA)

- The physical limitations on the rate of motion of the human vocal organ requires that human produces speech at an average rate of about 10 phonemes per second.

- If each phoneme is represented by a binary number, then a six bit numerical code (i.e., 64 possibilities) is sufficient to represent all of the phonemes of English.

- For an average rate of 10 phones per second and neglecting any correlation between pairs of adjacent phones (and assuming that all phones are equally probable) we get an estimate of $\underset{\text{6 bits}}{6} \times 10 =$ 60 bps for the average information rate of speech.

- Speech bandwidth is between 4 (telephone quality) and 8 kHz (宽频 wideband hi-fi speech). Consider sampling the speech signal at between 8 and 16 kHz, and using 8 (log-scale quantization) bits per sample.
  - 8000 × 8 = 64 kbps (telephone)
  - 16000 × 8 = 128 kbps (wideband)

sampling freq. must be twice as signal bandwidth

- This is between 1000 to 2000 times change in bit rate from discrete message symbols to digital representation of speech. To date, it is still a challenge to achieve three orders of magnitude of compression.

A brief note on entropy

- Let $A = \{s_0, s_1, .., s_{M-1}\}$ be the set of M possible symbols and $P(s_m)$ be the probability of occurrence of the symbol s_m. In the previous example, s_m being the phones. basic elements

- The expected information per symbol referred to as the entropy (a measure of the information content) is given by

$$\text{Entropy}, H = E\{-\log_2 P(s_m)\} = \sum_{m=0}^{M-1} [-\log_2 P(s_m)]P(s_m)$$

$$= -\sum_{m=0}^{M-1} P(s_m) \log_2 P(s_m)$$

- The unit for entropy is bit. The maximum entropy is achieved when all symbols in the set are equally probable (e.g., a fair dice, a fair coin, raining or sunny).

when doing FFT, always define like this

- Let $M = 2^B$ where B is an integer (e.g., the number of bits use to represent the discrete symbols), consider a uniform $P(s_m) = 1/M$, the entropy is given by

$$H = \log_2 M = B \log_2 2 = B$$

number of bits

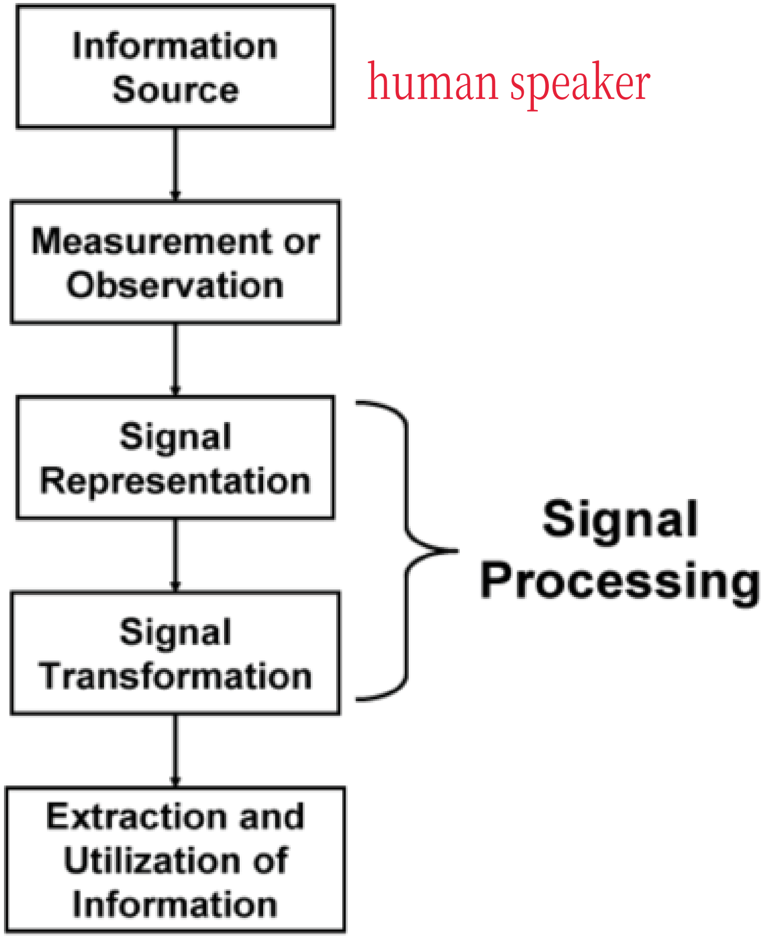
- The information rate is then given by the product of the entropy H and the signaling rate R. Given that $H = 6$ bits and $R = 10$ phones per second, the information rate is 60 bps.

M=2=2^1 so B=1, H=1 bit

- Consider a language with only two possible phones, the speaker might have to talk 6 times faster to achieve the same 60 bps.

Speech signal processing model

- Figure on the right shows a generic model of information processing from the speech message, where the human speaker being the information source, to the
  - human listener (e.g., in the case of speech enhancement, speech coding)
  - machine (e.g., in the case of speech and speaker recognition)



- In most speech applications, the measurement or observation is generally the acoustic waveform (the speech sounds). Other forms include the measurements of the positions (or neural control signals) of the articulators over time.
- Signal processing:
  - Representation of the signal based on a given model (e.g., spectral representation)
  - Transformation of signal to a more convenient form (e.g., spectral average across entire utterance)
- Extraction of information (e.g., speaker information, message)

Example: silent speech interface

- A device that allows speech communication without using the sound made when people vocalize their speech sounds, e.g., via electromyography (EMG) measurements of the neural control signals of the articulators over time.

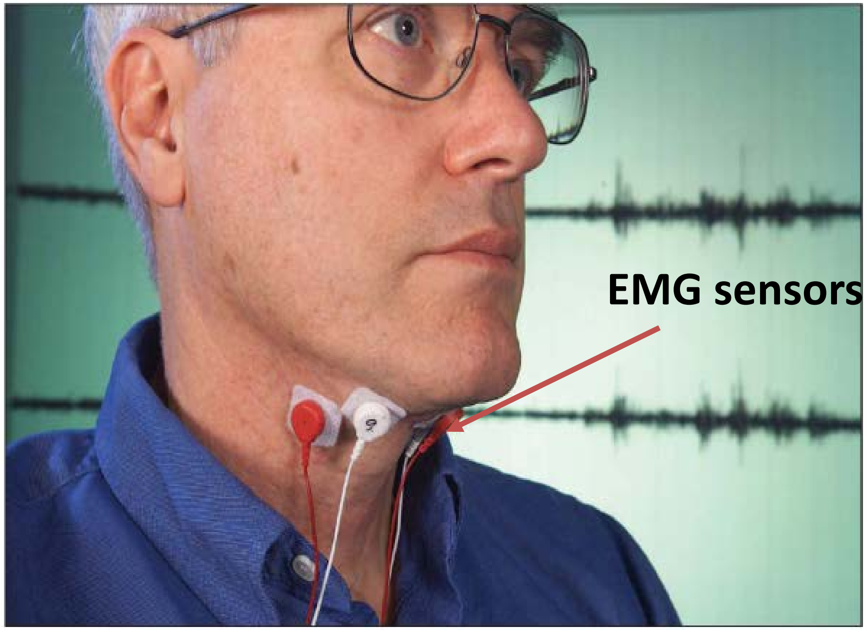


Figure 1. Stick-on sensors read silent signals that can then be processed by recognition software. (photo courtesy of NASA Ames Research Center, Dominic Hart)

[Source: L. McLaughlin, "Silencing voice recognition technology," IEEE Intelligent Systems, Vol. 19, no.3, pp. 4 - 6, 2004.]

EE6424 Part 2: Lecture 1.2

APPLICATIONS

Applications of digital speech processing

- The first step of most speech applications is to digitize the acoustic waveform, which essentially convert a continuous-time continuous-amplitude signal into a sequence of numbers with a finite resolution.

- Other representations (e.g., spectrogram, *mel-frequency cepstral coefficients*) are obtained by processing of the discrete-time representation.

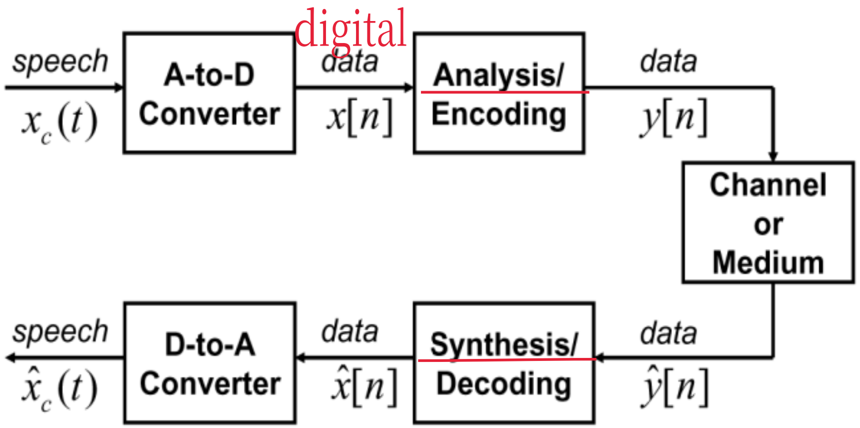
- Speech applications:
  - Speech coding and enhancement
  - Text-to-Speech synthesis 合成
  - Speech and speaker recognition
  - Speech-to-speech translation
  - Others

Speech coding

- The most widespread application of speech technology.
- The goal of a speech coder is to compress a digitized speech signal into a lower bit-rate representation while maintaining a desirable perceptual quality.
- The compressed representation allows efficient transmission and storage of speech signals for a broad range of applications including:
  - Wired telephony (landline), cellular communication, VoIP (which utilizes the internet as a real-time communication medium)
  - Telephone answering machine, interactive voice response (IVR)
- Speech coders often utilize many aspects of both speech production and perception of human
  - May not be useful for more general audio signals like music (source-filter model is not valid)

- General audio coders like MP3 and AAC incorporates only aspects of sound perception, and therefore do not achieve as much compression as speech coder (considering speech input).

- Figure below shows a block diagram of a generic speech coding system. The channel (or medium) indicates the transmission channel (or the storage medium).



- The transmitted $y(n)$ and received $\hat{y}(n)$ data can be identical with a properly designed error protection scheme.
- The coded and decoded signal $x(n)$ and $\hat{x}(n)$ are not exactly the same.

Recognition and pattern matching applications

- Automatic extraction of information from speech signal
  - Speech recognition – extract the message from the speech signal
  - Keyword spotting – monitoring a speech signal for the occurrences of specified words or phrases
- We speak to convey messages. The speech signal carries as well additional information of the speaker (gender, emotion, identity etc.)
  - Speaker recognition – identify who is speaking among a set of speakers
  - Speaker verification – verify a speaker’s claimed identity
  - Spoken language recognition – identify the language of a given utterance

Additional information are conveyed

Information type	Listen to the following samples		
Language	🔊	🔊	🔊
Gender	🔊	🔊	
Identity	🔊	🔊	

Speech-to-speech translation

发声

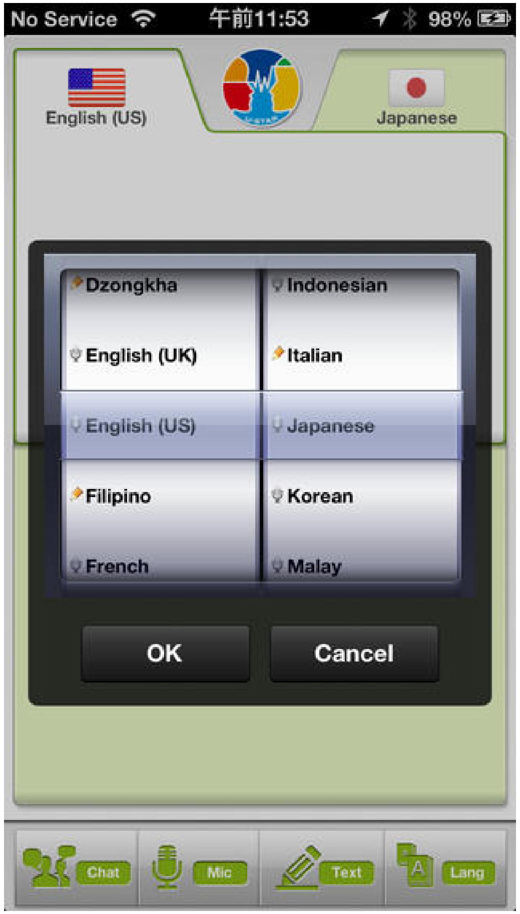
- Conversion of a spoken utterance in one language into another language, which may not have the same set of phonemes (i.e., they sound differently), e.g., English to Chinese and vice versa.
- Speech A → Message (Language A) → Message (Language B) → Speech B
- Involves the use of the following technologies:
  - Speech recognition
  - Text-to-text translation (e.g., Google translation)
  - Speech synthesis

中枢

- English is commonly use as the pivot language. For example:

  Vietnamese → (English) → Chinese.

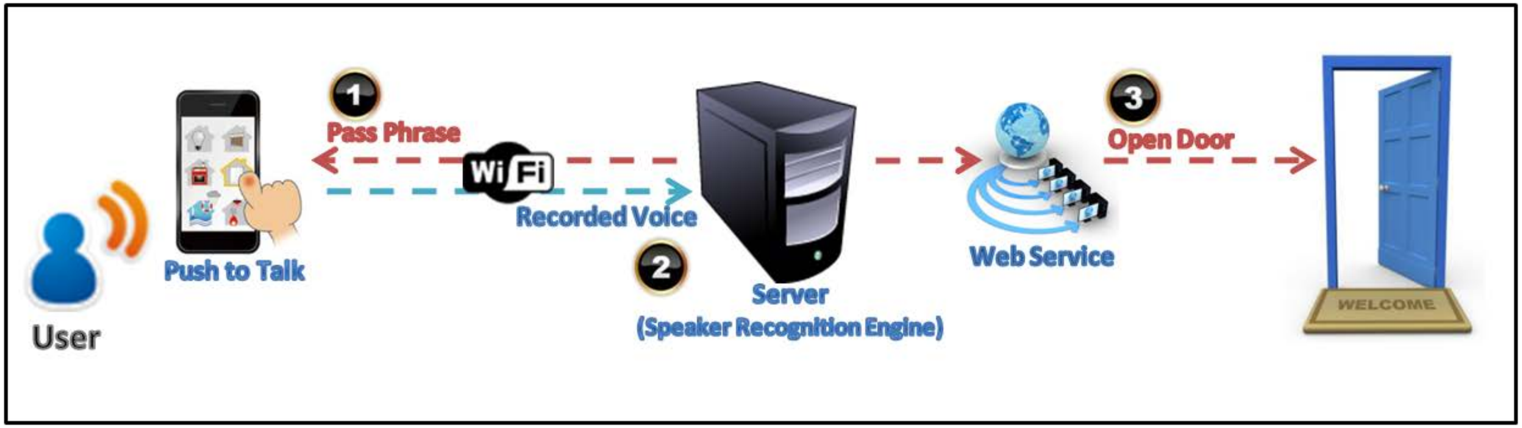
Overcoming language barriers around the world



- Developed by USTAR (Universal Speech Translation Advanced Research Consortium), an international research consortium for network-based speech-to-speech translation (S2ST) system.
- See http://www.ustar-consortium.com/ for more information.

STARHome@Fusionopolis

- **STARHome** is a fully functional 180 square meters smart home prototype located at the Fusionopolis, Singapore.
- Deployment of speaker verification for entrance-door access control and speech recognition (small vocabulary) for command control as part of home security and automation.
- Video (click here)



Summary

- The fundamental purpose of speech is communication, i.e., the transmission of messages (or ideas) from one person to the others.

- Speech is related to
  - Language (linguistics is a branch of social science)
  - Human physiological capability (physiology is a branch of medical science)
  - Sound and acoustics (acoustic is a branch of physical science)

- Purpose of speech processing:
  - To understand speech as a means of communication (Characteristics of Speech, Speech Analysis in the Time and Frequency Domains)
  - To represent speech for efficient transmission and reproduction (Sampling and Quantization of Speech, Speech Coding, Speech Enhancement)
  - To analyze speech for automatic recognition and extraction of information (Speech Recognition, Speaker Recognition, Spoken Language Recognition)