11-737 Multilingual NLP

Speech

Carnegie Mellon University
Language Technologies Institute

Table of Contents

- What is speech?
- Speech applications
- Speech databases
- Speech hierarchy

What is speech???

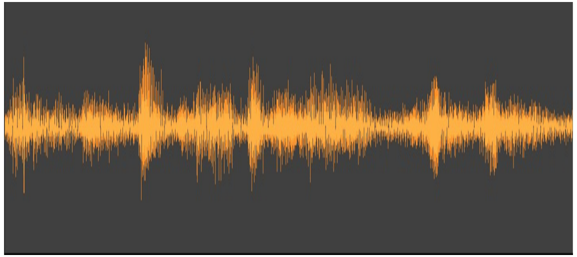
Watanabe’s definition

- **Sound produced by human for the communication**
- Is this speech?

- Freesound https://freesound.org/

Sound

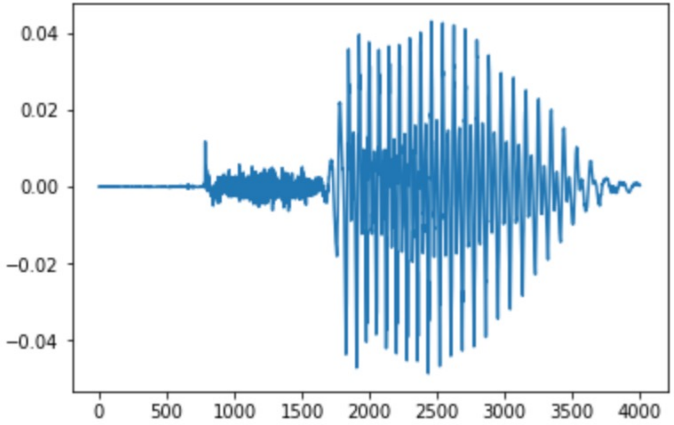
- Air pressure
- Captured by a microphone



- Governed by well known physical properties
  - Attenuation, refraction, reflection, diffraction, superposition

Speech waveform?

- Waveform: Converting a sound pressure into a time series
- Usually ***1-dimensional waveform*** (mono) in this lecture
  - A lot of recording devices support stereo waveforms.
  - Then, it would be 2 (left and right) dimensional waveform
  - We also use a microphone array to capture N-dimensional waveform where N means the number of channels captured by N microphones (e.g., Alexa has 7 microphones, N=7)



What kind of information does speech sound contain?

- Transcription
- Speaker identity
- …
- …

Table of Contents

- What is speech?
- Speech applications
- Speech databases
- Speech hierarchy

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Speech…
- Spoken…

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

Any others?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

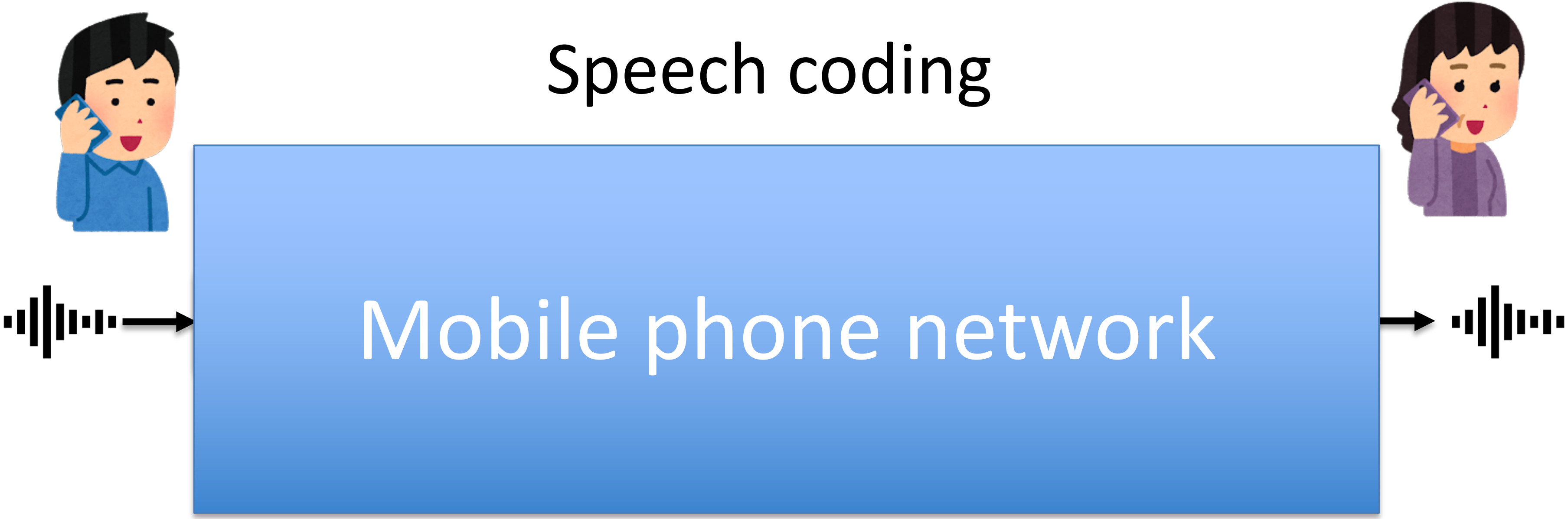
What is the **most widely** used technique among them?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

What is the **most widely** used technique among them?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- **Speech coding**
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

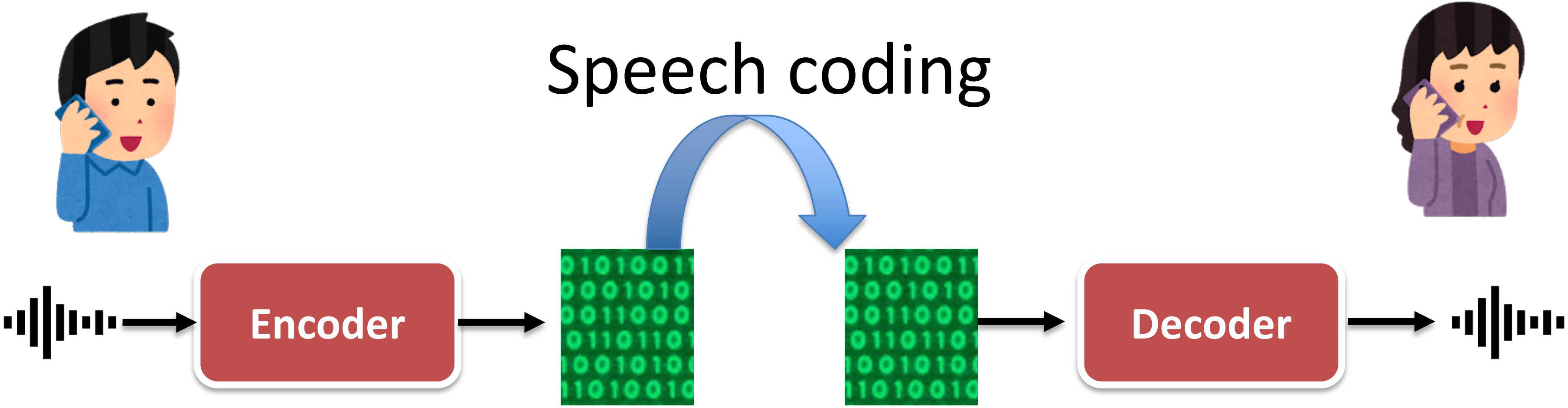
Speech coding



Mobile phone network

Our infrastructure

Speech coding



Our infrastructure

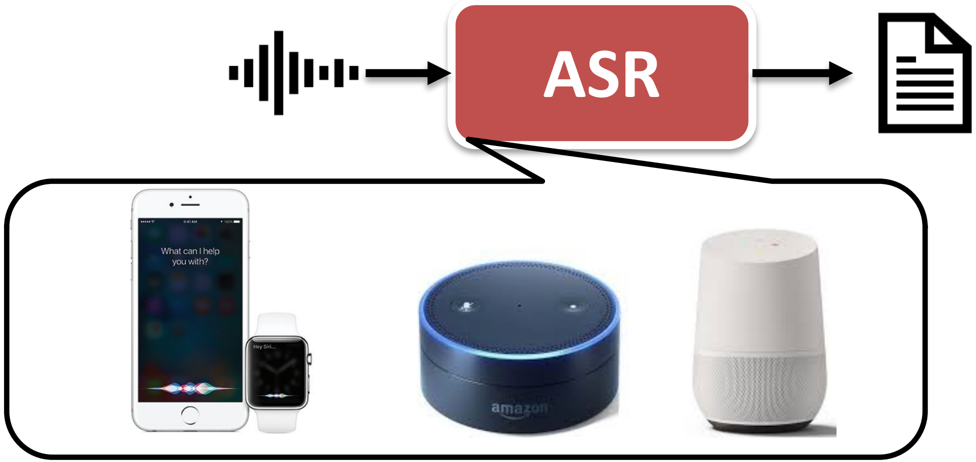
Compress the data while keep the speech information

Statistical method based on linear prediction

What kind of research topics in speech research?

- **Speech recognition**
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

Automatic Speech Recognition (ASR)



ASR

Widely used in many applications!

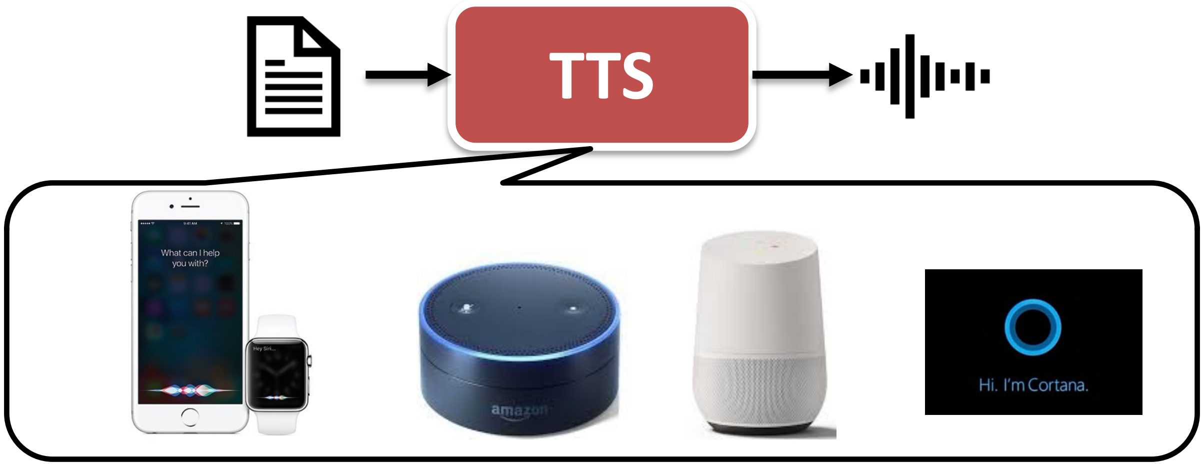
We will discuss it in more details in the next lecture

What kind of research topics in speech research?

- Speech recognition
- **Speech synthesis**
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception

- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

Speech Synthesis
(TTS: Text to Speech)



TTS

Inverse problem of ASR

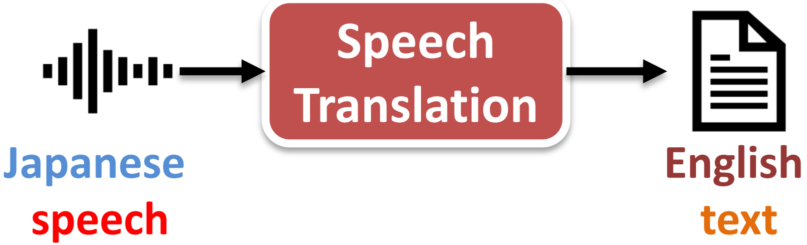
We will discuss it week after the next week

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception

- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- **Speech translation**
- Multimodal processing
- Speech corpus

Speech Translation
source speech to target text

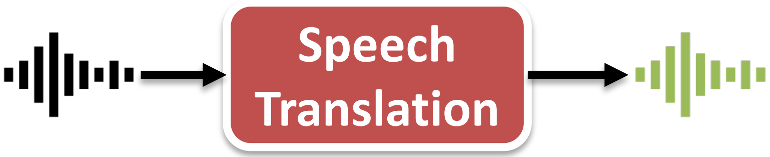


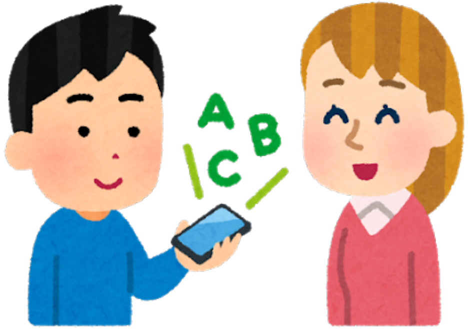
Combining ASR + machine translation

☹Complicated systems, Error Propagation

End-to-End modeling has been actively studied

Speech Translation
source speech to target speech





ASR + machine translation + TTS

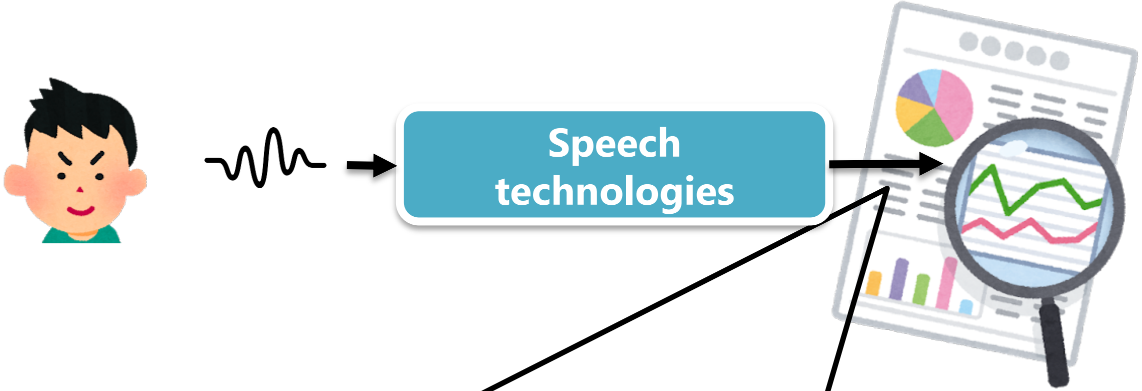
End-to-end?

One of the goals of multilingual NLP

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Voice conversion
- **Speaker recognition**
- **Language recognition**
- **Speech emotion recognition**
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- **Audio event classification and detection**
- Speech separation
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

Speaker Profiling/Audio Disentanglement



Speech technologies

Linguistic information

Speaker Characteristics, gender, age

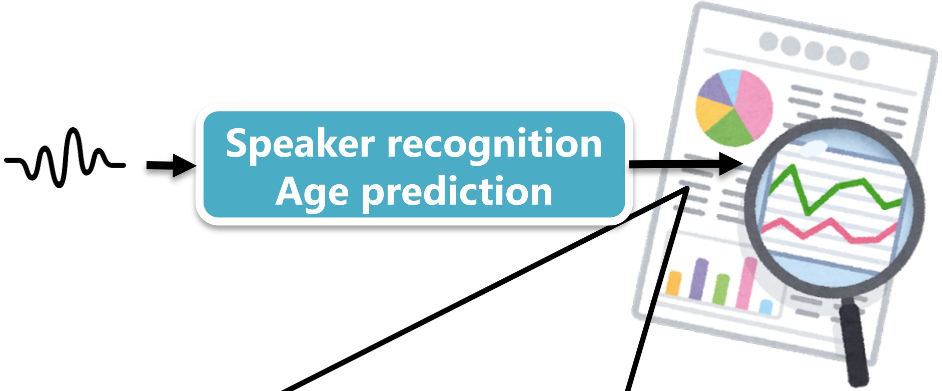
Language

Emotion

Audio events

Etc,

Speaker Profiling/Audio Disentanglement



Speaker recognition
Age prediction

Linguistic information

Speaker Characteristics, gender, age

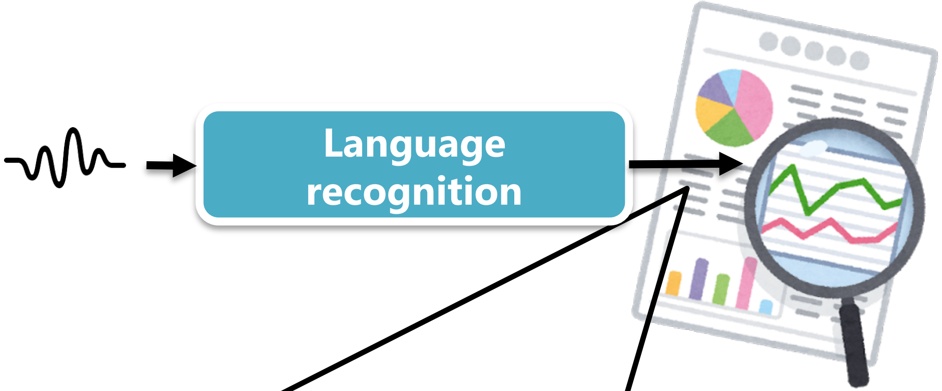
Language

Emotion

Audio events

Etc,

Speaker Profiling/Audio Disentanglement



Language recognition

Linguistic information

Speaker Characteristics, gender, age

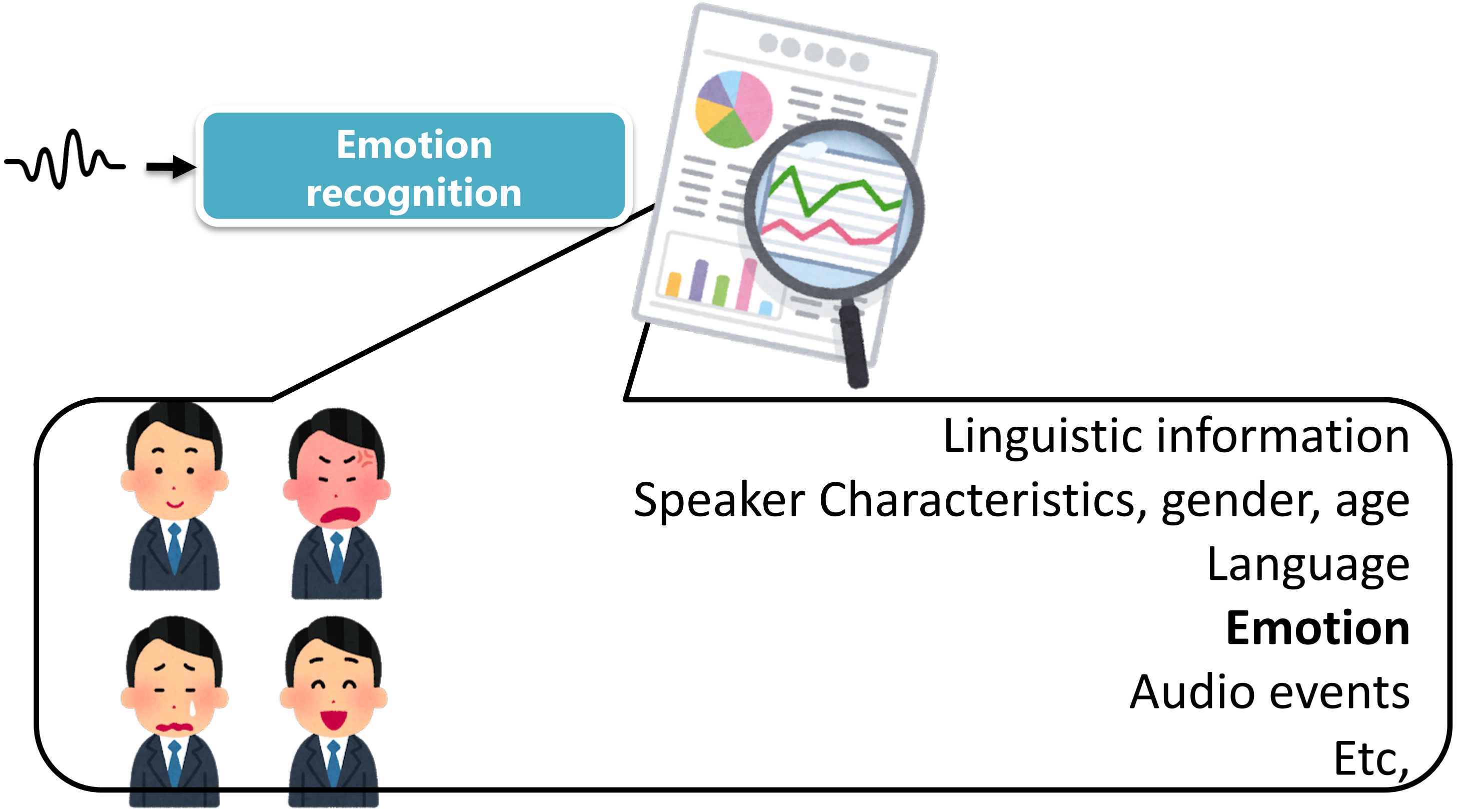
Language

Emotion

Audio events

Etc,

Speaker Profiling/Audio Disentanglement



Emotion recognition

Linguistic information

Speaker Characteristics, gender, age

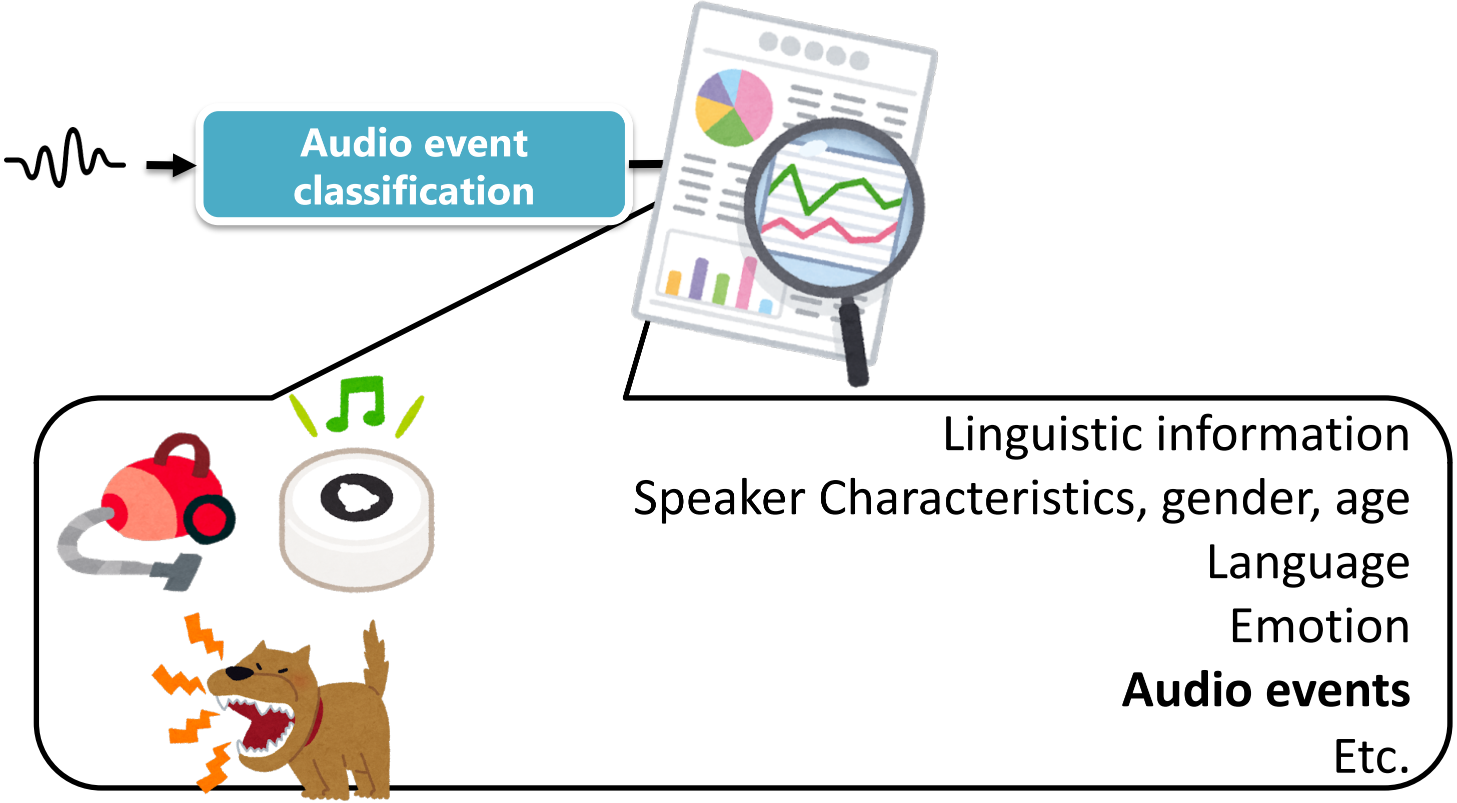
Language

Emotion

Audio events

Etc,

Speaker Profiling/Audio Disentanglement



Audio event classification

Linguistic information

Speaker Characteristics, gender, age

Language

Emotion

Audio events

Etc.

Privacy in speech

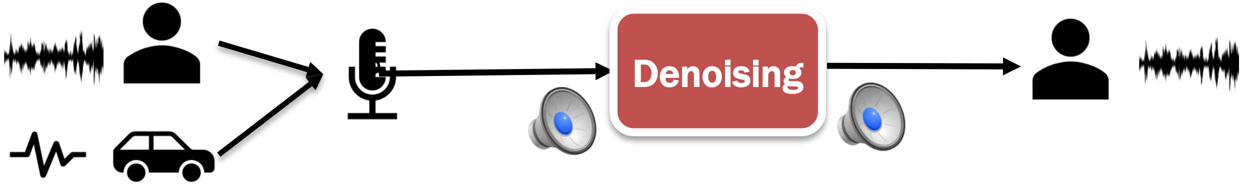
- Speech contains various profiling information
- Current speech processing techniques require massive computations
  - Most computations at a server
  - Serious privacy issues
  - On device AI

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- **Speech enhancement**
- Microphone array processing
- Audio event classification and detection
- **Speech separation**
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

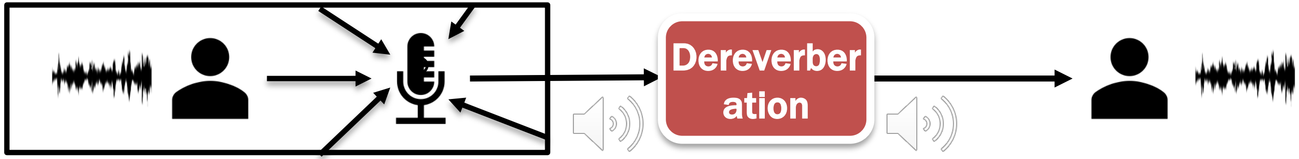
Speech enhancement
Several types of problems

- Denoising (people mainly call it speech enhancement)



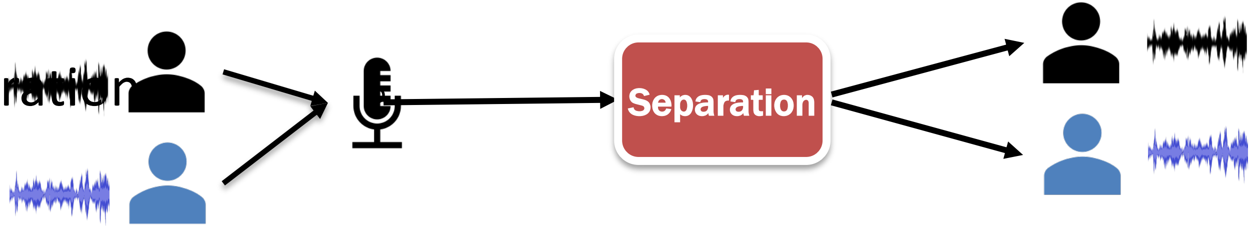
Denoising

- Dereverberation



Dereverberation

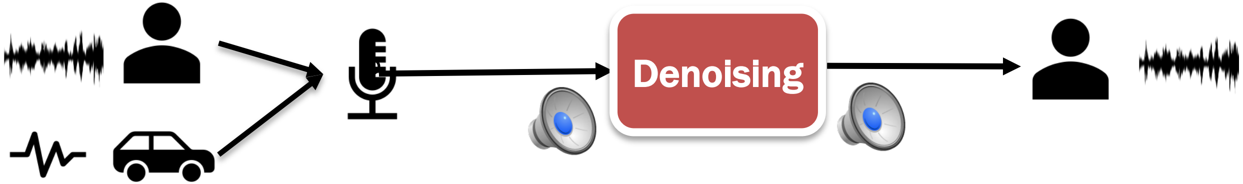
- Separation



Separation

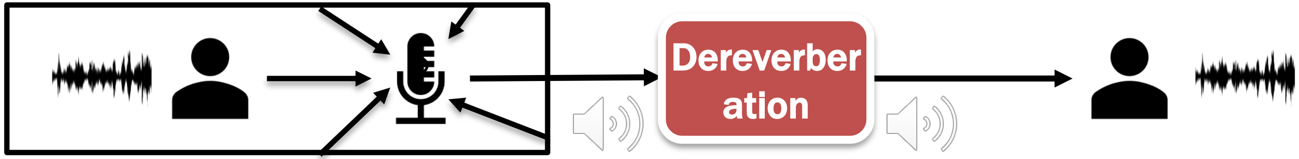
Speech enhancement
Several types of problems

- Denoising (people mainly call it speech enhancement)



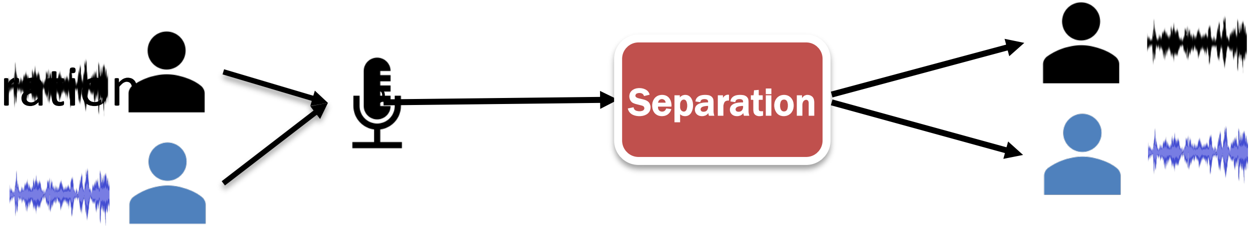
Speech enhancement
Several types of problems

- Dereverberation

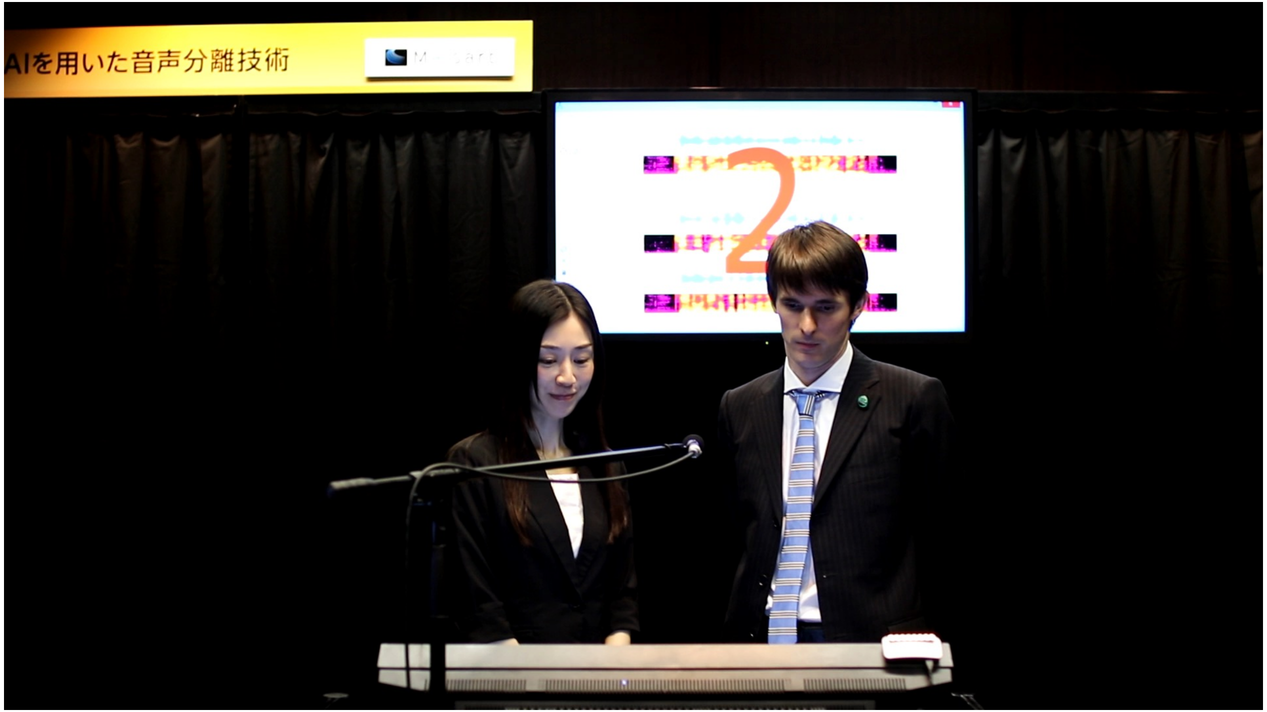


Speech enhancement
Several types of problems

- Separation



Deep clustering based speech separation [Hershey et al., 2016]



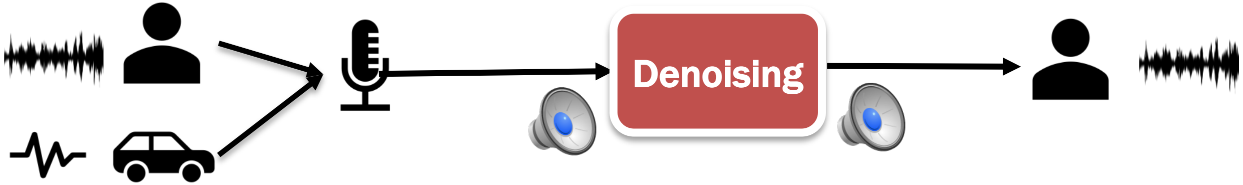
How many microphones do we have?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception

- **Speech enhancement**
- **Microphone array processing**
- Audio event classification and detection
- **Speech separation**
- Spoken language understanding
- Spoken dialogue systems
- Speech translation
- Multimodal processing
- Speech corpus

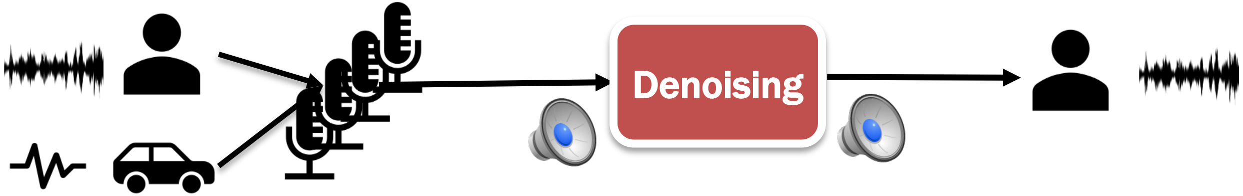
Microphone array processing
Single to multiple microphones

- Denoising (people mainly call it speech enhancement)



Microphone array processing
Single to multiple microphones

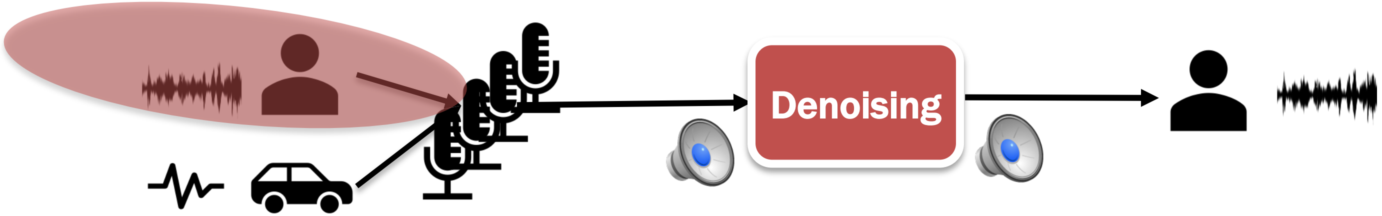
- Denoising (people mainly call it speech enhancement)



Microphone array processing

Single to multiple microphones

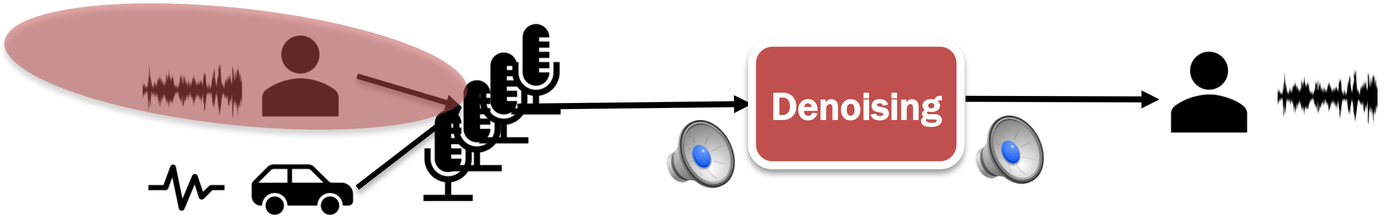
- Denoising (people mainly call it speech enhancement)



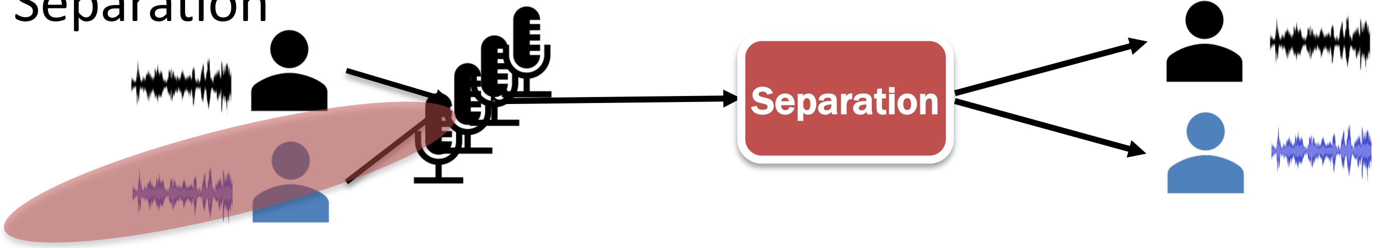
Make a spatial **beam** (beamforming)
to only pick up desired signals

Microphone array processing
Single to multiple microphones

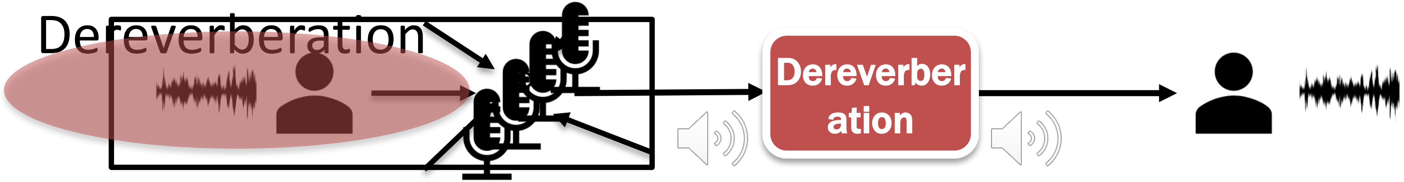
- Denoising (people mainly call it speech enhancement)



- Separation



- Dereverberation



Cocktail party

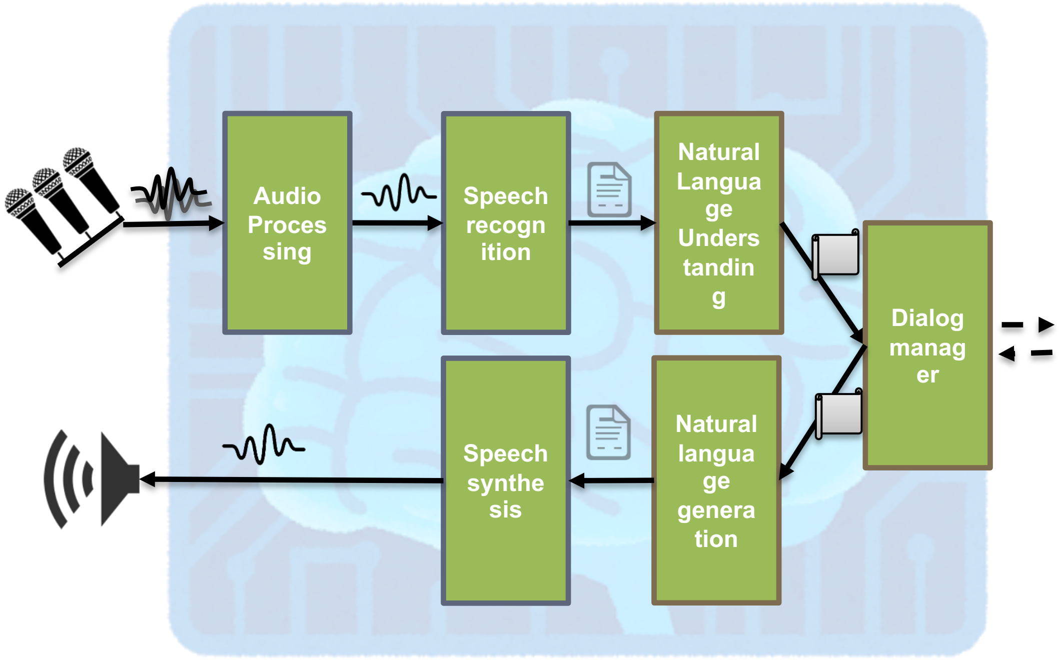
- Many systems have more than one mic.
  - Alexa 7
  - Human 2
  - More microphones, easier to listen
- **Cocktail party**
  - Human can easily understand
  - One of the most difficult problem for a machine
- **One of the important speech research goal is to realize “who is speaking when what where how”**

What kind of research topics in speech research?

- Speech recognition
- Speech synthesis
- Voice conversion
- Speaker recognition
- Language recognition
- Speech emotion recognition
- Speaker diarization
- Speech coding
- Speech perception
- Speech enhancement
- Microphone array processing
- Audio event classification and detection
- Speech separation
- Spoken language understanding
- **Spoken dialogue systems**
- Speech translation
- Multimodal processing
- Speech corpus

My long-term research topic
Conversational AI

Spoken dialog systems



Audio Processing

Speech recognition

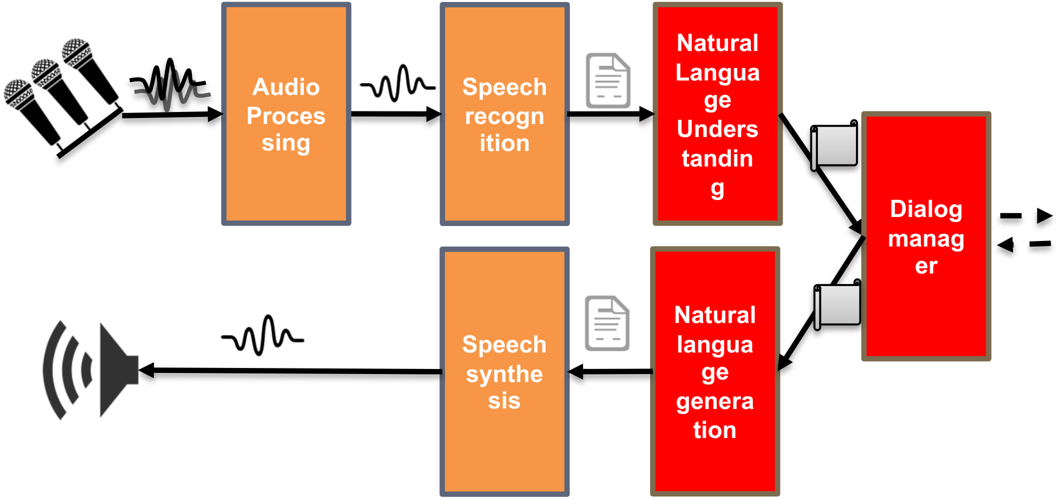
Natural Language Understanding

Dialog manager

Natural language generation

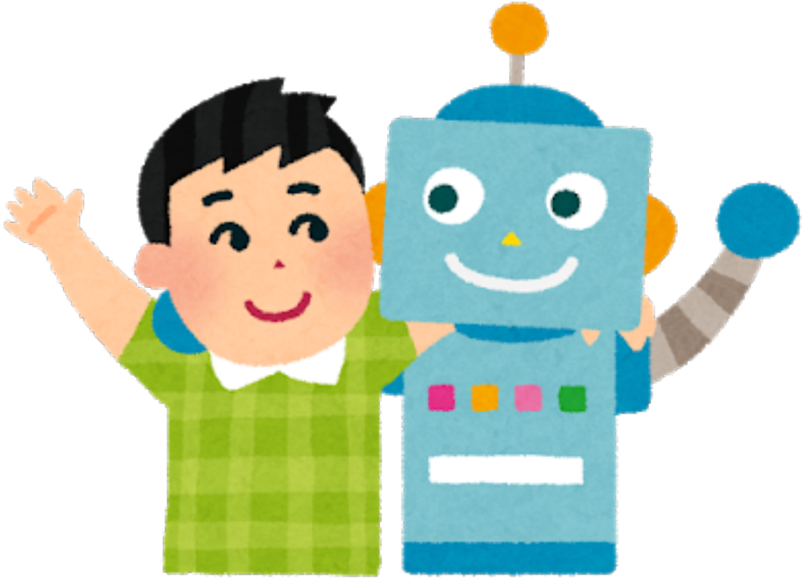
Speech synthesis

Speech + Language!



One of the ultimate speech research goals

- Human-level spoken dialog systems



What kind of research topics in speech research?

- **Speech recognition**
- **Speech synthesis**
- **Voice conversion**
- **Speaker recognition**
- **Language recognition**
- **Speech emotion recognition**
- **Speaker diarization**
- **Speech coding**
- **Speech perception**
- **Speech enhancement**
- **Microphone array processing**
- **Audio event classification and detection**
- **Speech separation**
- **Spoken language understanding**
- **Spoken dialogue systems**
- **Speech translation**
- **Multimodal processing**
- **Speech corpus**

Table of Contents

- What is speech?
- Speech applications
- Speech databases
- Speech hierarchy

Speech variations
Speaking styles and environments

	Style	Hours	Environment	Transcriber
Wall Street Journal (WSJ)	Read speech	~80	Clean/Close talk	Just confirm
Switchboard	Spontaneous	~300	Clean/Close talk	Have to transcribe
Librispeech	Read speech	~1,000	Clean/Close talk	Just confirm
CHiME-3	Read Speech	~20	Noisy/Distant talk	Just confirm
CHiME-6	Spontaneous	~50	Noisy/Distant talk	Have to transcribe

- Read speech: we prepare sentences in advance, and ask people to read them
    - Easy to obtain the reference
- Non-read speech (spontaneous): we have to transcribe by listening the audio, expensive

Read speech examples

- Read a prompt
- We can make a pair data of a prompt and corresponding audio

Ex) common voice: https://commonvoice.mozilla.org/en

- Easy to collect
  - We still need to check whether the person can correctly utter a prompt
- Easy to anonymize
- Not a real conversation

Spontaneous speech

- Transcribe actual recording

- Real, real, real
- Takes very long time to transcribe it
  - 2 minutes of the switchboard audio sample takes 30 minutes (for the beginner)
  - Need some postprocessing (anonymization, filler handling, etc.)

Single speaker processing to conversation processing

Single speaker

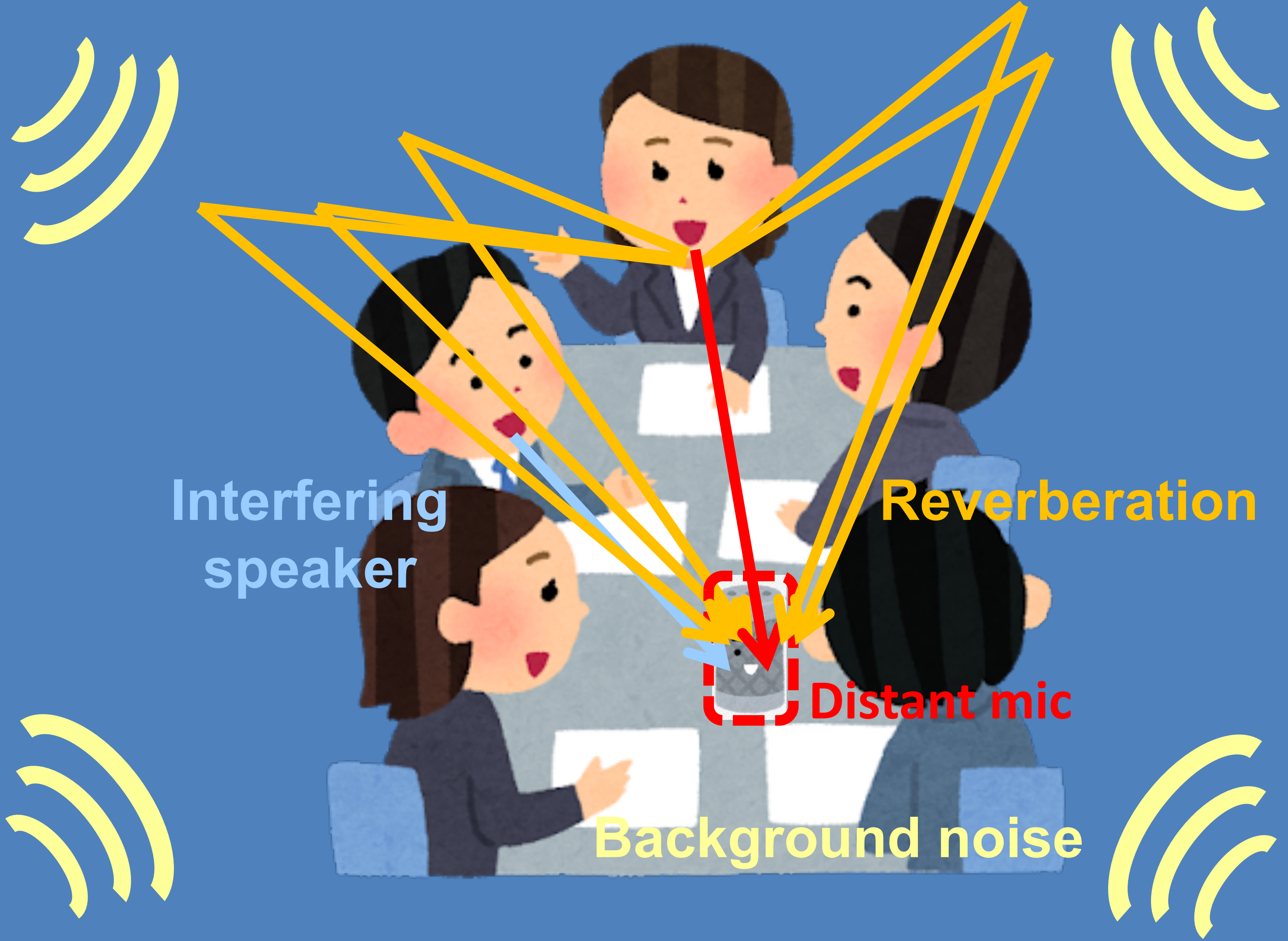
Close-talking microphone

Error rate <5 %

Conversation analysis

Distant microphone

Error rate ~40%



Interfering speaker

Reverberation

Distant mic

Background noise

CHiME-3

http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/

🔊 Cafe

🔊 Street

🔊 Bus

🔊 Pedestrian area

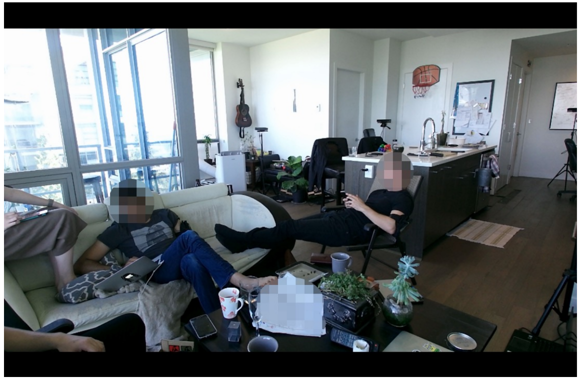
CHiME-6 examples

https://chimechallenge.github.io/chime6/

🔊

🔊

🔊

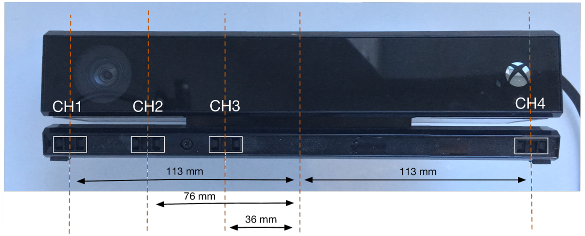


🔊

The CHiME-6 recording setup

Data has been captured with 32 audio channels and 6 video channels

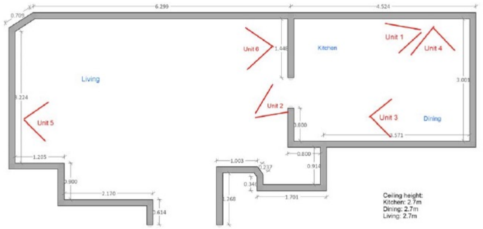
- Participants' microphones
  - Binaural in-ear microphones recorded onto stereo digital recorders
  - Primarily for transcription but also uniquely interesting data
  - Channels: 4 x 2
- Distant microphones
  - Six separate Microsoft Kinect devices
  - Two Kinects per living area (kitchen, dining, sitting)
  - Arranged so that video captures most of the living space
  - Channel: 6 x 4 audio and 6 video



Example recording setups

S04

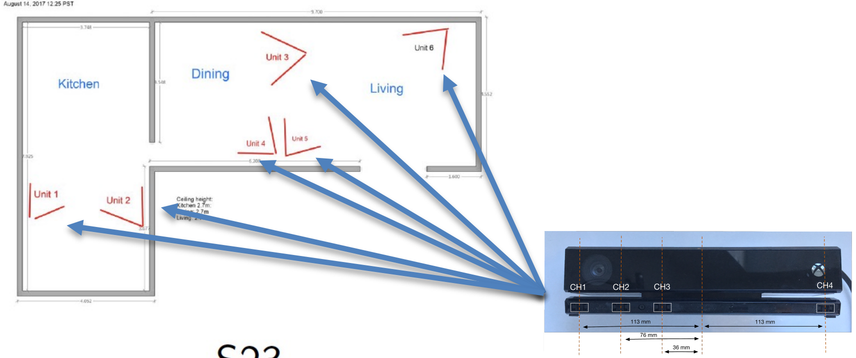
Session ID: S04
August 12, 2017 18:03 PST



Ceiling height:
Kitchen: 2.7m
Dining: 2.7m
Living: 2.7m

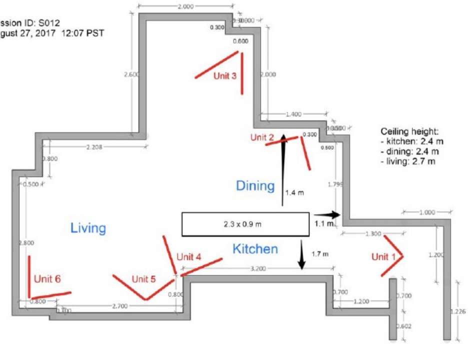
S07

Session ID: S07
August 14, 2017 12:25 PST



S12

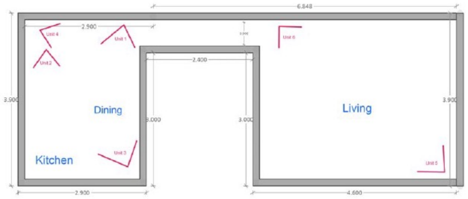
Session ID: S012
August 27, 2017 12:07 PST



Ceiling height:
- kitchen: 2.4 m
- dining: 2.4 m
- living: 2.7 m

S23

Session ID: S23
September 22, 2017 14:53 PST

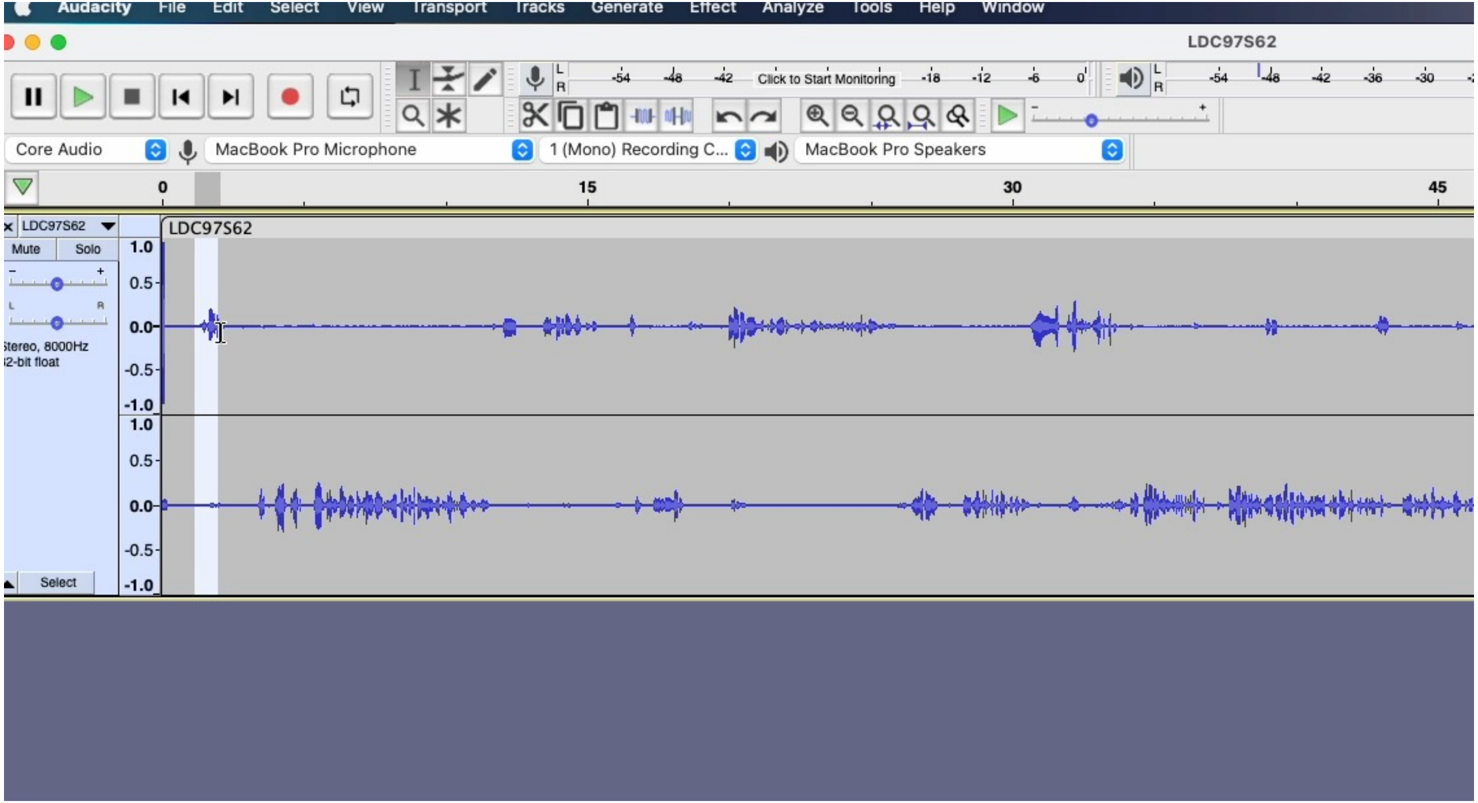


Ceiling height:
- kitchen: 2.9 m
- dining: 2.9 m
- living: 3.3 m

Spontaneous speech

- Transcribe actual recording
  - Example based on Audacity (developed at CMU!)
- Real, real, real
- Takes very long time to transcribe it
  - 2 minutes of the switchboard audio sample takes 30 minutes (for the beginner)
- Need some postprocessing (anonymization, filler handling, etc.)

How to transcribe an audio with Audacity?



Where we found the speech data?

- LDC, ELRA, other university or government institution
  - https://www.ldc.upenn.edu/
  - Well managed, license restricted
  - Famous ASR benchmarks (e.g., TIMIT, WSJ, Switchboard)
- Voxforge, openslr, commonvoice, zenodo
  - We can find less license restricted data (e.g., Creative Commons)
- Audio books, public recordings with captions (e.g., YouTube, Podcast, TED talk, Parliament or other government recordings, Bible)
  - Need some cares for the license and post processing
  - The data will be updated very frequently (deletion, modification, API change, etc.)
  - CMU Wilderness has **700(!)** languages (20 hours each)

How many hours of training data do we need?

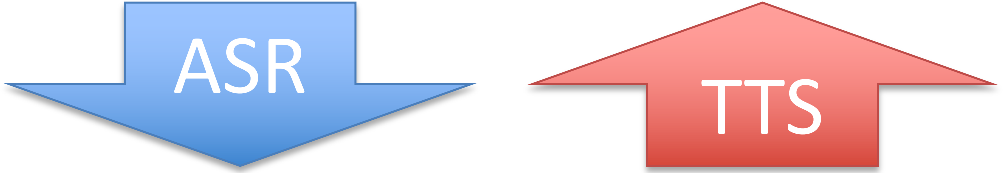
- We often use "**hour**" as a unit
- Commercial products: **More than thousand hours**
  - Very limited languages as public data, e.g., English, Mandarin, Japanese, German, Russian
- Do some ASR research experiments: **~100 hours**
- Less than 100 hours: Low-resource language in ASR
  - Pre-training/fine-tuning is changing the game

Table of Contents

- What is speech?
- Speech applications
- Speech databases
- Speech hierarchy

Speech <-> Text

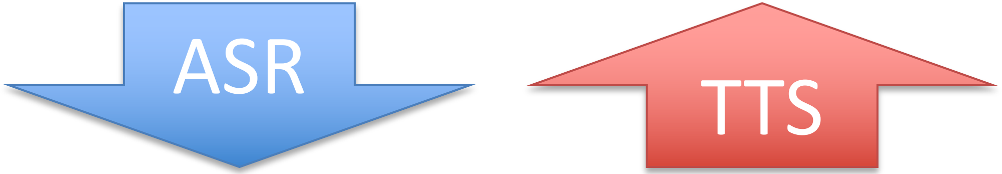
Speech sound: 🔊



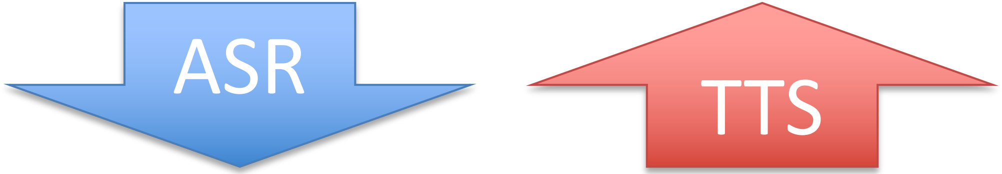
Text: I want to go to the CMU campus

Speech <-> Phonem <-> Text

Speech sound:



Phoneme: AY W AA N T T UW G OW T UW DH AH S IY EH M Y UW K AE M P AH S



Text: I want to go to the CMU campus

What is phone and phoneme???
GO TO: “g oʊ t u” or “G OW T UW”

- Phone: g oʊ t u
  - Devised by International Phonetic Association
  - Not applicable to all languages, needs special characters, too many variations, use of them depending on linguists

- Phoneme: one of the units of that distinguish one word from another in a **particular language**
  - /r/ and /l/ are degenerated in some languages (e.g., “rice” and “lice” sounds same for me!)
  - ARPAbet vs. International Phonetic Alphabet (IPA)
  - ARPAbet: `G OW T UW`
    - Proposed by ARPA for the development of speech recognition of only “American English”
    - Represented by ASCII characters

Pronunciation dictionary

- CMU dictionary
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict

”I want to go to the CMU campus”

→AY W AA N T T UW G OW T UW DH AH S IY EH M Y UW K AE M P AH S

- Powerful, but limited
- Out of vocabulary issue, especially new word

→ Grapheme2Phoneme mapping based on machine learning

Let’s play the CMU dictionary!

- Access: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Find some in-vocabulary words
- Find five out-of-vocabulary words

Multilingual phone dictionary

- https://en.wiktionary.org/wiki/Wiktionary:Main_Page

Multilingual speech recognition (phone based)

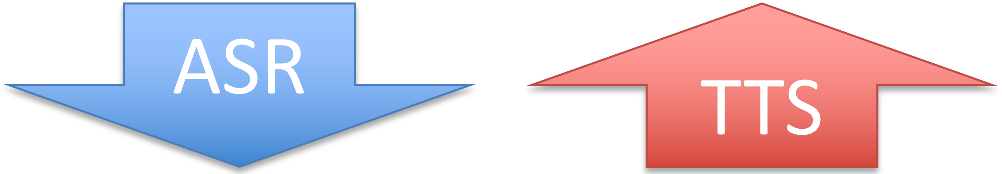
- Try to split the problem from speech to phoneme and phoneme to text
- Speech to phone**: language independent (acoustic model)**
- Phone to phoneme, phoneme to word: **language dependent (lexicon model)**

→

- Build speech to phone based on universal acoustic model
- Linguistic knowledge to make a lexicon model

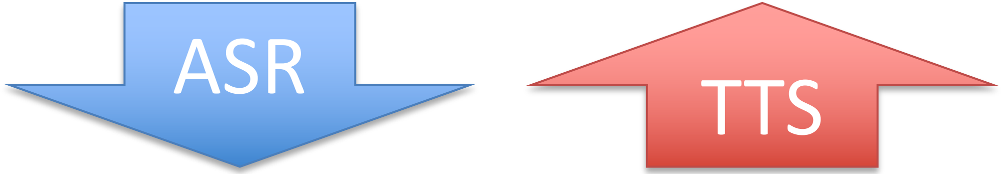
Speech <-> Phonem <-> Text

Speech sound:



Language **independent**

Phoneme: AY W AA N T T UW G OW T UW DH AH S IY EH M Y UW K AE M P AH S



Text: I want to go to the CMU campus

Language **dependent**

Multilingual speech recognition (phone based)

- Try to split the problem from speech to phoneme and phoneme to text
- Speech to phone**: language independent (acoustic model)**
- Phone to phoneme, phoneme to word: **language dependent (lexicon model)**

→

- Build speech to phone based on universal acoustic model
- Linguistic knowledge to make a lexicon model

Other units?

- Syllable {C*} V {C*}
- Allophone: /k/ can be different depending on the context (/a/-)/k/(-**/a/**), (/a/-)/k/(-**/i/**)
- Pinyin
- Etc.

Summary of today’s talk

- Speech: sound waveform but used by human for the communication
- Speech applications: many applications
- Speech data: read vs. spontaneous, various sources
- Speech hierarchy: introduction of phone and phoneme

- The next lectures will introduce two main applications, ASR and TTS

Assignment 3