CS11-737 Multilingual NLP

Automatic Speech Recognition

Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

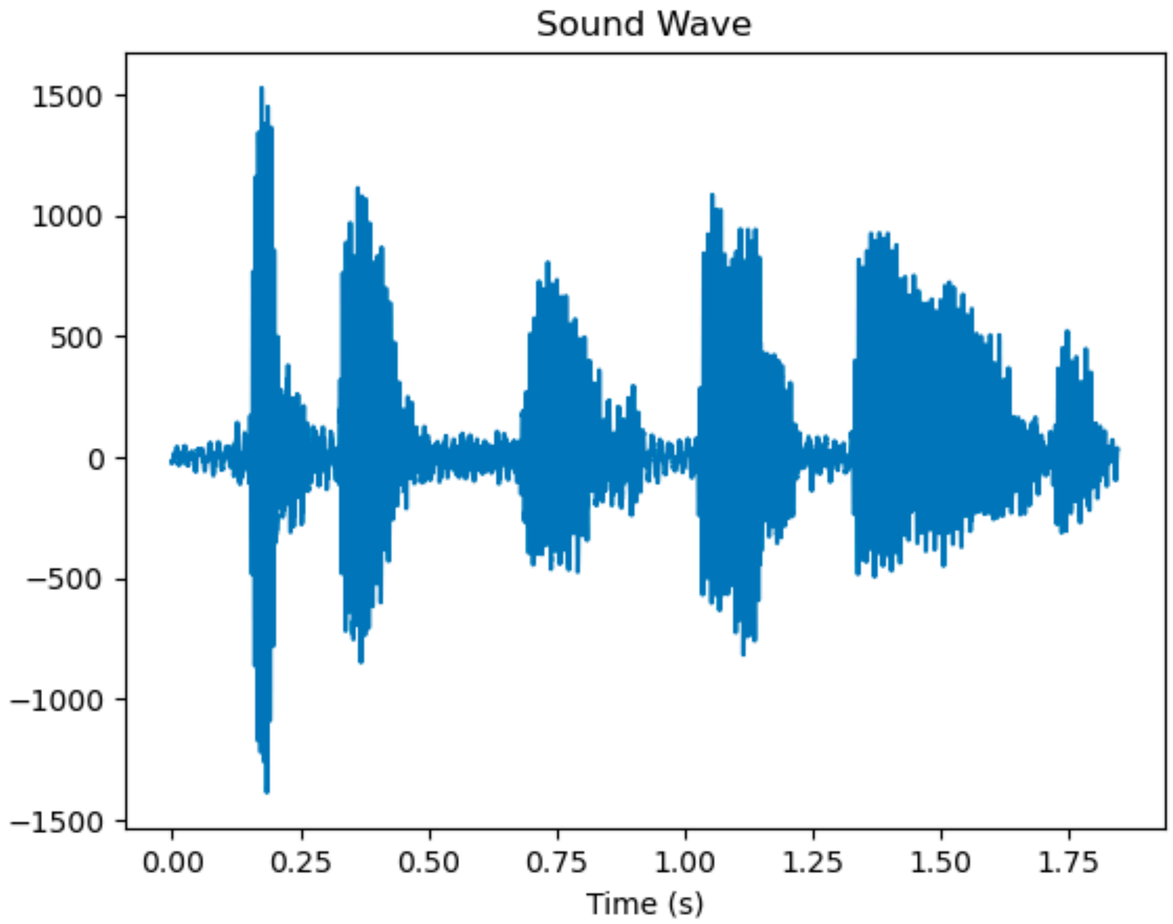
Carnegie Mellon University
Language Technologies Institute

What is speech

- Sound produced by human for communication
- versus noise, or other non-human sound

Speech Data

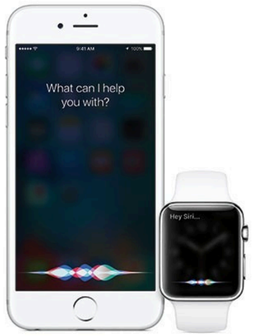
- Sound by air pressure
- Recorded by microphone
  - mono: recorded by a single microphone
- Waveform: a time series of recorded air pressure



Speech Processing Tasks

- Speech Recognition
  - speech to text transcript
- Speech synthesis
  - text to speech
- Language identification
- Speech separation
- Speech attribute classification
- Voice cloning
- Speech translation

Automatic Speech Recognition (ASR)



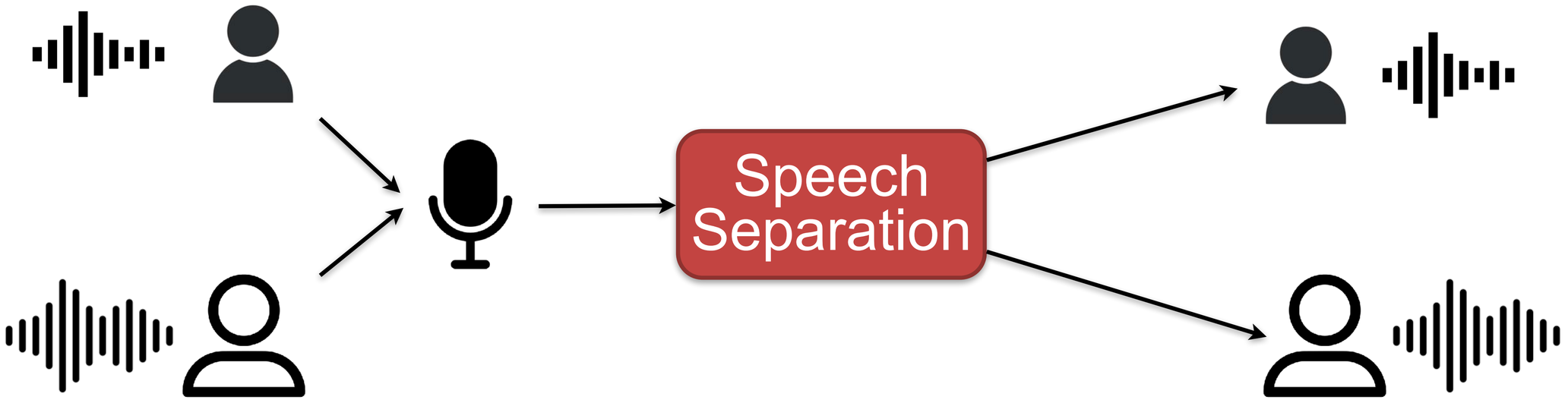
Sphinx, Janus, ESPnet, etc.
Developed and
maintained By CMU!

Widely used in many applications!

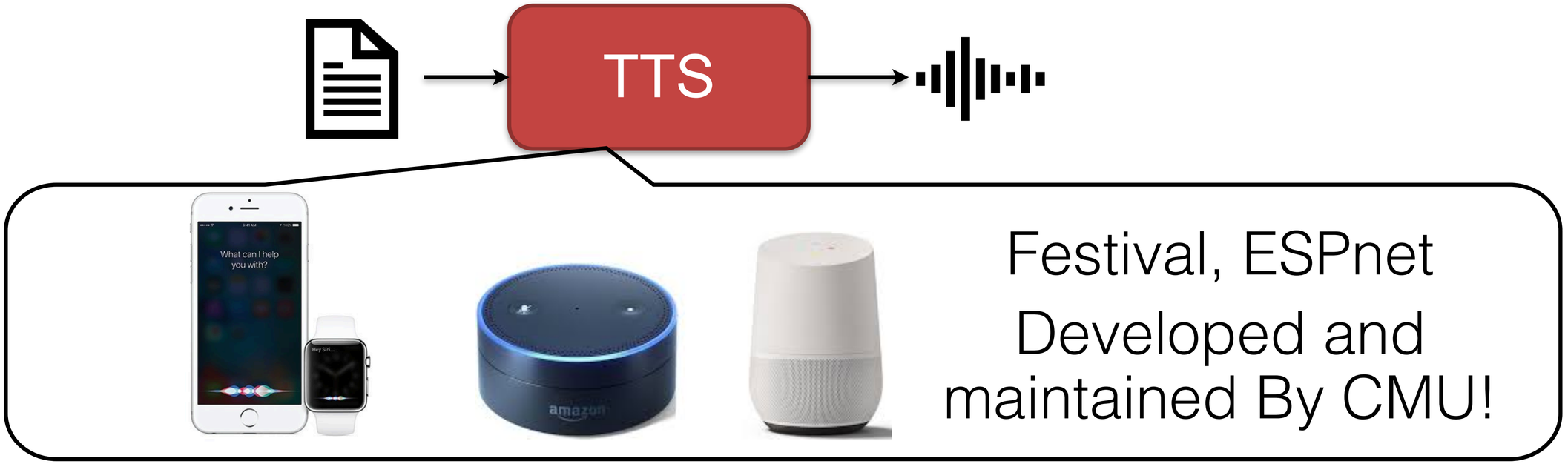
Why is ASR hard?

- noise
- speed
- different words with same prounnunciation
- accent
- distance to mic
- dialect
- multi-person interaction
- Spoken vs written, filler words

Speech Separation



Speech Synthesis (Text to Speech, TTS)



Inverse problem of ASR

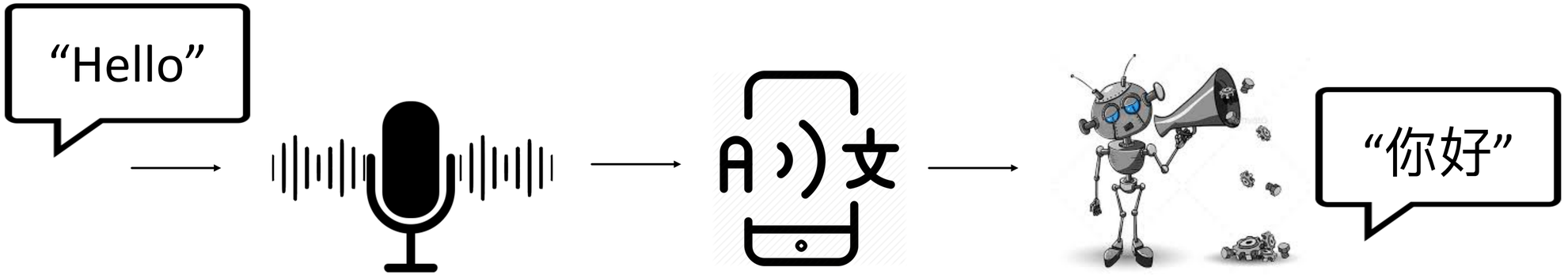
TTS with Voice Cloning

	baseline	VALL-E
They moved thereafter cautiously about the hut groping before and about them to find something to show that Warrenton had fulfilled his mission.		
Her husband was very concerned that it might be fatal.		

Speech Translation (ST)

- source language ***speech(audio)*** → target lang ***text***

- source language ***speech*** → target lang ***speech***



Ultimate goal is crossing all language barriers in human communication

Speech Recognition at Cocktail Party

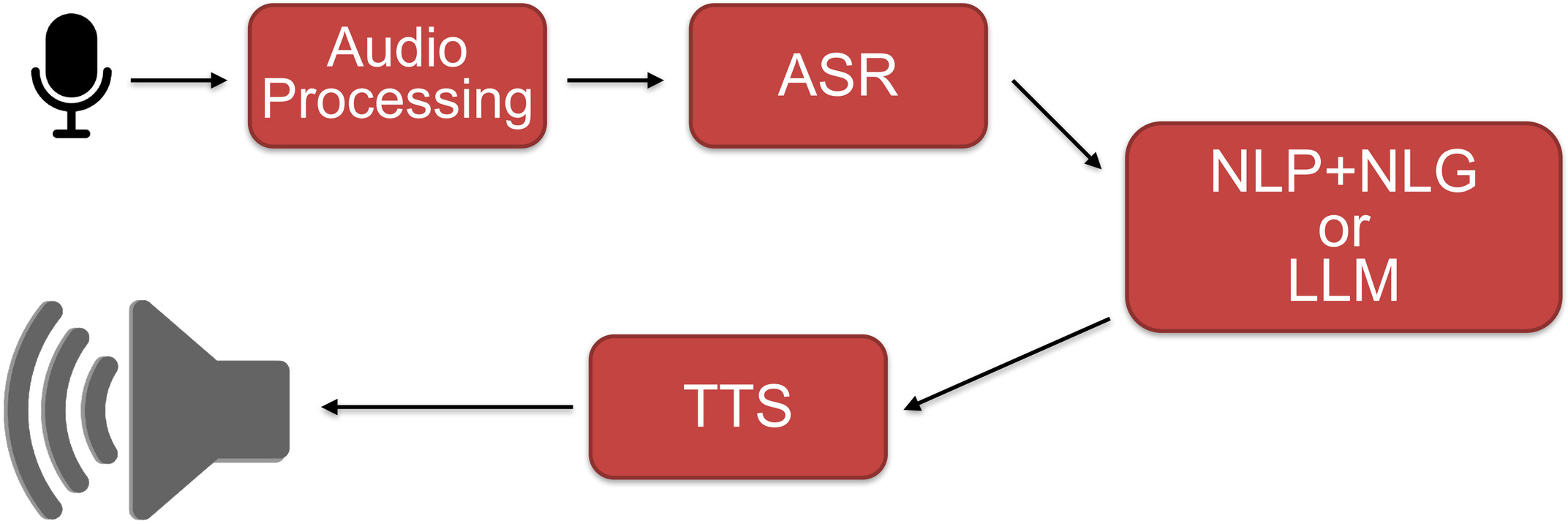
- Many systems have more than one mic
  - Alexa: 7
  - Human: 2

- Human can easily separate and understand conversations at cocktail party

- Big challenge: speech understanding to capture "who is speaking what when where how"

Speech + Language Processing

Spoken dialog system



Type of Speech Data

- Read speech: record reading of text (reference given)
- Non-read speech (spontaneous): have to transcribe the audio, expensive

	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
Common Voice	Read speech, 114 languages	3,300		

Common Voice

https://commonvoice.mozilla.org/

Validated Hours

19,160

Recorded Hours

28,751

Languages

114

What's inside the Common Voice dataset?

Each entry in the dataset consists of a unique MP3 and corresponding text file. Many of the **28,751** recorded hours in the dataset also include demographic metadata like age, sex, and accent that can help train the accuracy of speech recognition engines.

The dataset currently consists of **19,160** validated hours in **114** languages, but we're always adding more voices and languages. Take a look at our Languages page to request a language or start contributing.

Spontaneous Speech

- Transcribed actual recording

- Huge cost to transcribe
  - 1 minutes of the switchboard audio sample takes 30 minutes
  - need postprocessing (anonymization, filler handing, etc)

Single Speaker versus Conversation

- Single speaker
- close-talking microphone
- ASR error rate < 5%

- Conversation analysis
- distant microphone
- Error rate ~ 40%

CHiME-6 Challenge: https://chimechallenge.github.io/chime6/overview.html

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

Speech Data

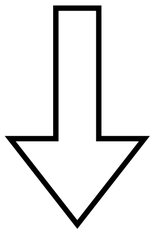
- LDC: www.ldc.upenn.edu
- Voxforge, openslr, commonvoice, zenodo
  - often less restricted license
- Audio books, public recordings with captions
  - Youtube, Podcast, TED talk, Parliament recordings, bible
  - care the license
  - CMU Wilderness has 700(!) languages (20 hours each)

How much data is needed?

- Commercial ASR product: thousands of hours
- ASR research: ~ 100 hours
- Low-resource ASR: < 100 hours
  - pre-training/finetuning

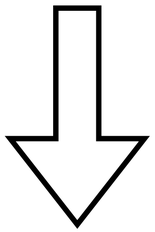
From Speech to Text

Phoneme



P IH T S B ER G . IH Z . AH . S IH T IY . AH V . B R IH JH .

Transcript text



Pittsburgh is a city of bridge

Phoneme

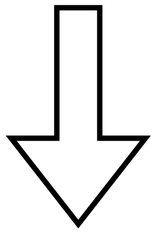
- Phone:
  - unit of sound in speech, regardless meaning
  - International Phonetic Alphabet (IPA)
  - not applicable to all languages

- Phoneme:
  - a set of sounds that can distinguish one word from another
  - /sɪn/ vs. /sɪŋ/
  - /pæt/ vs. /bæt/ vs. /ˈbɛt/ vs. /bʌt/
  - /θɪk/ vs. /sɪk/

Pronunciation Dictionary

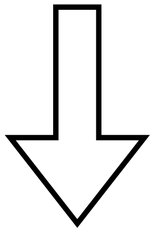
- CMU dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- 134k words in American English
- Out-of-vocabulary (especially new words)
  - Grapheme2Phoneme: LOGIOS
  - grapheme is the smallest unit of a writing system for a language

Pronunciation for Chinese



Pinyin pǐ zī bǎo shì qiáo liáng zhī dū

IPA $p^{hj}i\text{ː}^{3}ts\text{ɻ̩ː}^{1}pau^{3}\ \text{ʂ}\text{ɻ̩ː}^{4}\ t\text{ɕ}^{h}au^{2}l^{j}a\text{ŋ}^{2}\ \text{ʈ}\text{ʂ}\text{ɻ̩ː}^{1}\ t^{w}u\text{ː}^{1}$



Transcript text

匹兹堡 是 桥梁 之 都

Pittsburgh is bridge of city

Pronunciation Difference across Languages

- Chinese phoneme represented by "**x-**", "**q-**", and "**j-**" in pinyin do not exist in English.

谢 (thanks)	xiè	/ɕi̯ɛ/
七(seven)	qī	/t͡ɕʰi⁵⁵/
巨(huge)	jù	/t͡ɕy⁵¹/

- Sound in English but not much difference for (some part) of Chinese people: e.g. /l/ and /n/

fault => faunt

Multilingual phone dictionary

https://en.wiktionary.org/

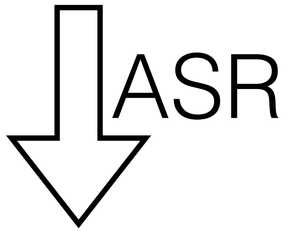
What are sounds in your language not in English?

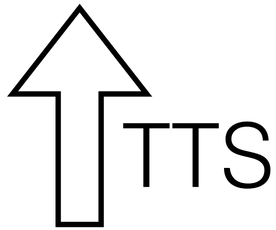
Multilingual Speech Recognition (phone-based)

- Split the problems: speech-to-phone 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 (e.g. language model)

Speech <-> Phoneme <-> Text

Phoneme

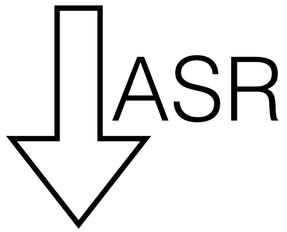


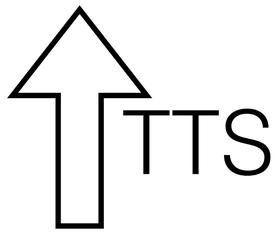


Language independent

P IH T S B ER G . IH Z . AH . S IH T IY . AH V . B R IH JH .

Transcript text





Language dependent

Pittsburgh is a city of bridge

Impact on Multilingual Speech Processing

- Pre-training based purely on raw audio data
  - pre-trained models can be applied to down-stream tasks or even a different language
  - XLSR: Unsupervised Cross-lingual Representation Learning for Speech Recognition
- Cross-lingual transfer

Summary

- Speech: sound waveform used by human for communication of information
- Speech technology: ASR, TTS, ST, etc
- Speech data: read vs. spontaneous
- Speech hierarchy: phone and phoneme, language difference
- Next lecture: ASR

Announcement

- Office hour of Lei Li: at GHC 5417 every Tuesday 4-5pm (instead of GHC 6403)

- Late policy: 3 total late days allowed without penalty, beyond that 10% penalty for each late day
  - Project final report does not have late day (need to turn in final grade)

Discussion

- Project ideas