Introduction and overview

HSE - IU Speech Recognition Workshop

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

What is speech recognition?

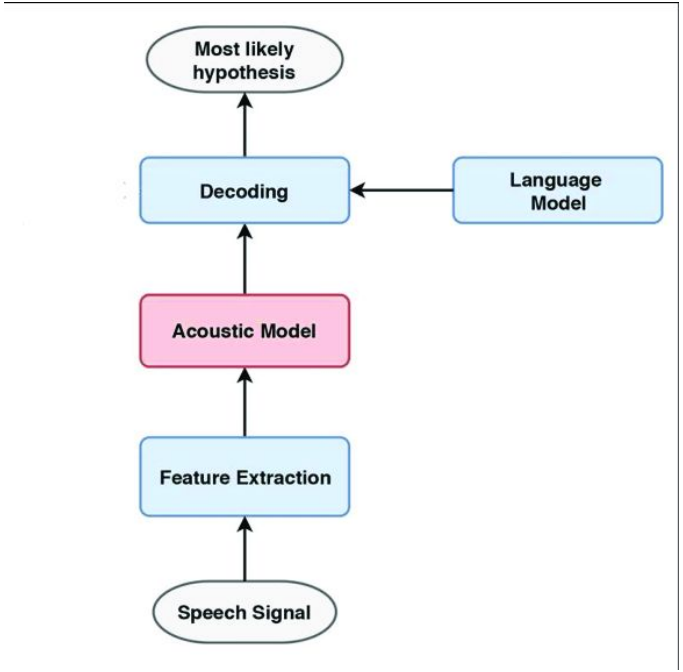
- A computer system that turns audible speech into readable text
- Also known as:
  - speech to text (STT), automatic speech recognition (ASR), …

Applications

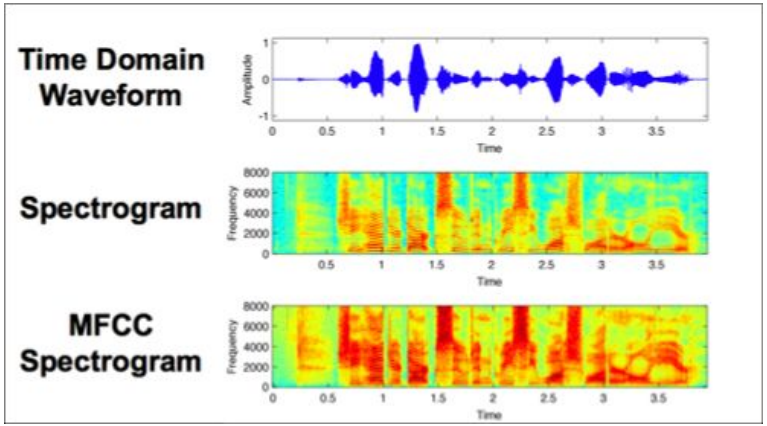
- Automatic subtitling and dictation software
- Voice control in cars, home assistants, etc.
- Pronunciation training for L2 speakers
  - https://papareo.nz/docs/PapaReo_NeurIPS2020_Poster.pdf

Components

- Audio processing (or feature extraction):
  - Convert audio into a machine-readable and processable form
- Acoustic model:
  - Take sequences of audio frames and predict how likely each character in the alphabet is for that frame
- Decoder:
  - Take a sequence of predictions and produce a set of hypotheses
- Language model:
  - Include information about spelling and context to improve predictions



Audio processing

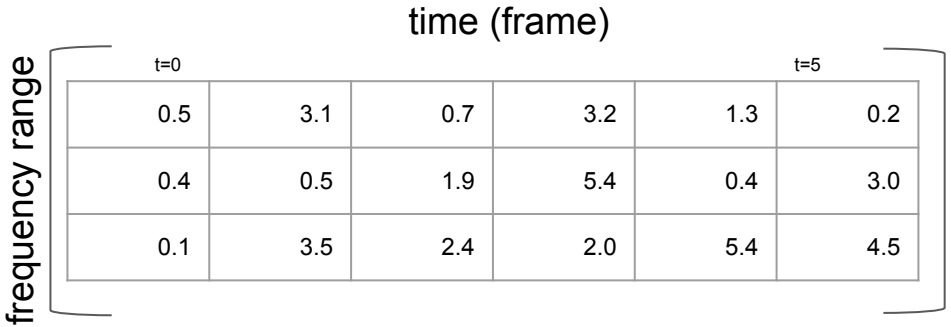


0.4	3.0	5.0	2.3	0.9	0.7

0.5	3.1	0.7	3.2	1.3	0.2
0.4	0.5	1.9	5.4	0.4	3.0
0.1	3.5	2.4	2.0	5.4	4.5

- Turn audio files into a machine-readable form (matrices of numbers)
- Go from **amplitude** over time to amplitude at frequency over time (2D → 3D)
  - Note: The higher the amplitude the higher the energy (loudness)

Acoustic model



alphabet

time (frame)

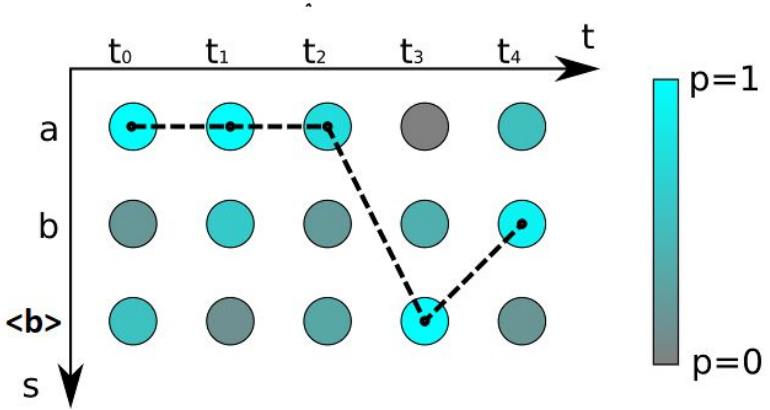
	a	b	c	...	k	...	t	...	z
t=0	0.1	0.1	**0.3**	...	**0.4**	...	0.1	...	0.0
	0.0	0.1	**0.3**	...	**0.5**	...	0.1	...	0.0
	0.7	0.0	0.0	...	0.1	...	0.1	...	0.1
	0.7	0.0	0.0	...	0.2	...	0.1	...	0.1
	0.0	0.1	0.2	...	0.1	...	**0.6**	...	0.0
t=5	0.0	0.0	0.1	...	0.1	...	**0.7**	...	0.1

- Predicts for each frame the **distribution** over the **alphabet**
  - Alphabet is the set of possible output characters [a, b, c, …, z]
- The best prediction here would be {k k a a t t}
  - Note: There are many more audio frames than alphabetic characters

Decoder

time (frame)

	a	b	c	...	k	...	t	...	z
t=0	0.1	0.1	**0.3**	...	**0.4**	...	0.1	...	0.0
	0.0	0.1	**0.3**	...	**0.5**	...	0.1	...	0.0
	0.7	0.0	0.0	...	0.1	...	0.1	...	0.1
	0.7	0.0	0.0	...	0.2	...	0.1	...	0.1
	0.0	0.1	0.2	...	0.1	...	**0.6**	...	0.0
t=5	0.0	0.0	0.1	...	0.1	...	**0.7**	...	0.1



- Takes predictions from the acoustic model and outputs final hypotheses
- Solves two problems:
  - Multiple frames can correspond to a single character
  - Finding the most probable sequence of characters given the input frames

Language modelling

{c a t, k a t, c a t t, k a t t}

w	*P(w)*
c a t	0.8
k a t	0.1
c a t t	0.05
k a t t	0.05

- There is typically far less transcribed audio data than raw text data
- Language models are trained on raw text and provide additional information:
  - Spelling in context, e.g. /tu/ = {to, two, too}
  - Word segmentation, e.g. /aisorit/ = [i, saw, it]

Remainder of the workshop

- 11:00 - 11:15 Introductions and installation
- 11:15 - 11:30 1. Introduction and overview
- **11:30 - 12:15 2. Audio processing (Anurag)**
- 12:15 - 12:30 Discussion and break
- **12:30 - 13:15 3. Acoustic models (Nastya)**
- **13:15 - 14:00 4. Decoding (Fran)**
- 14:00 - 14:15 Discussion and break
- **14:15 - 15:00 5. Language modelling (Nils)**
- 15:00 - 15:15 Wrapping up