Automatic Speech Recognition

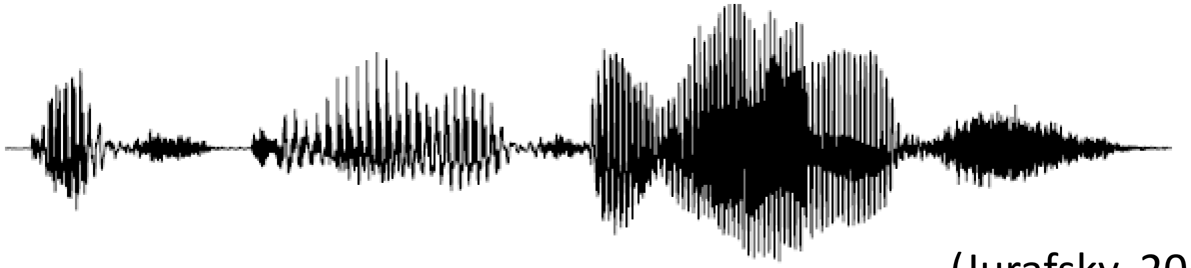
Sumire Honda

Overview

- What is ASR?
- Use Case
- History
- ASR Task
- Typical Architecture
- Evaluation
- Problem and Challenge
- Future

What is ASR?

- Definition
  - The conversion of speech into text (Foteini, 2020)
  - Map any waveform to the appropriate string of words (Jurafsky, 2020)



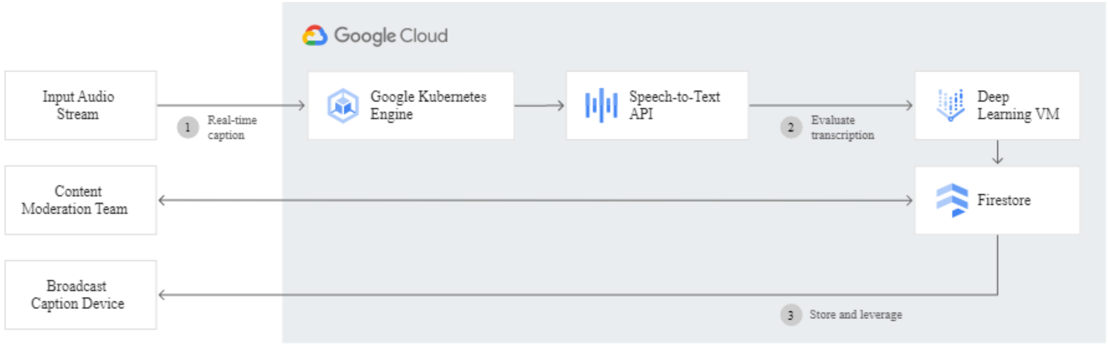
(Jurafsky, 2020)

It's time for lunch!

Use Case

- Human-machine communication for personal use
  - Communicating with smart home appliances
  - Generating captions for audio or video text

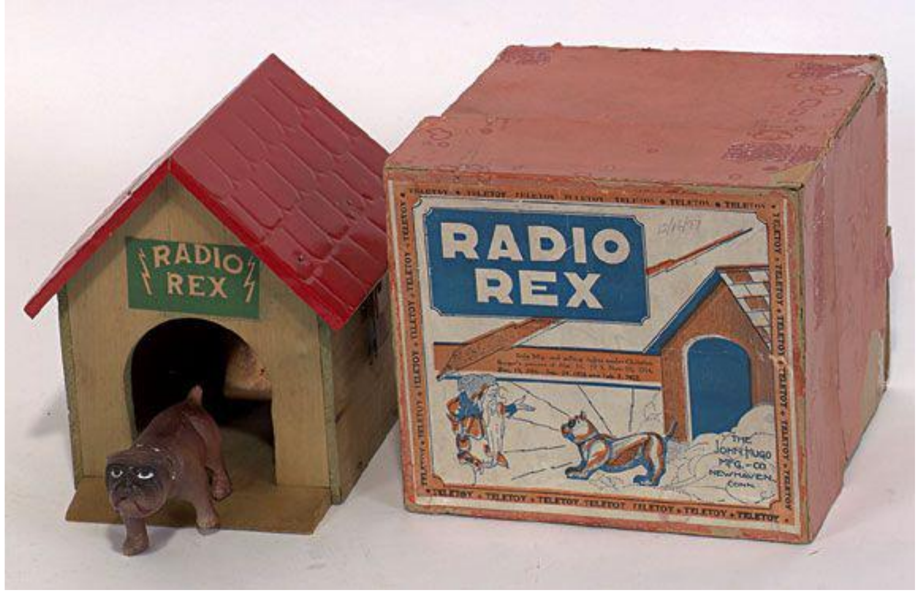
- Industrial Use
  - Industrial machine guidance with voice commands
  - Automatic telephone communication
  - Communication with automotive systems
  - Military vehicles
  - Communication with health care
  - Aerospace



(Google, use case for transcribing multimedia content)

History –The first “Speech Recognition”

- “Radio Rex” (1920s)
  - The first machine that recognize speech
  - Dog “Rex” comes by the spring was released by 500 Hz acoustic energy by the vowel [eh] in “Rex”



(Jurafsky, 2020)

2’ 55 https://youtu.be/AdUi_St-BdM?t=175

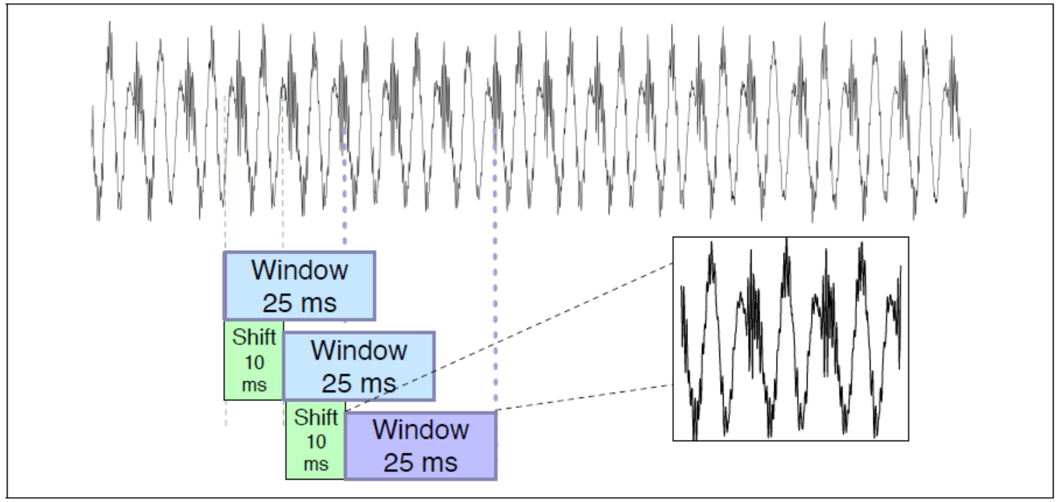
ASR Task

Breaks down waveform into very small window, where it represents a phoneme

Convert the analog waveform representations into a digital signal

Transform the digital signal into a sequence of acoustic feature vectors

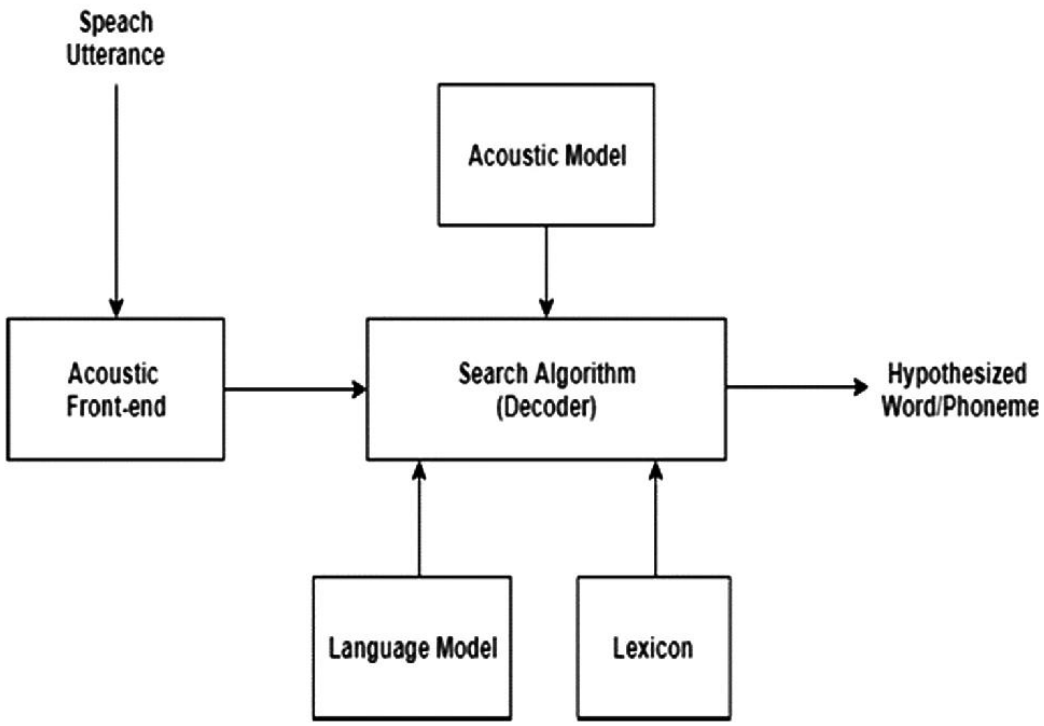
Use Algorithm to find the most probable word fit in that language.



(Jurafsky, 2020)

Architecture

- Acoustic front-end
  - Extract useful features from speech
- Language Model
  - Gives the limitation of the sequence of the given words
- Lexicon
  - Includes vocabulary
- Search Algorithm (Decoder)
  - Produce the hypothesized word/phoneme
- Acoustic Model
  - Contains Statistical representation of each sounds that makes up a word

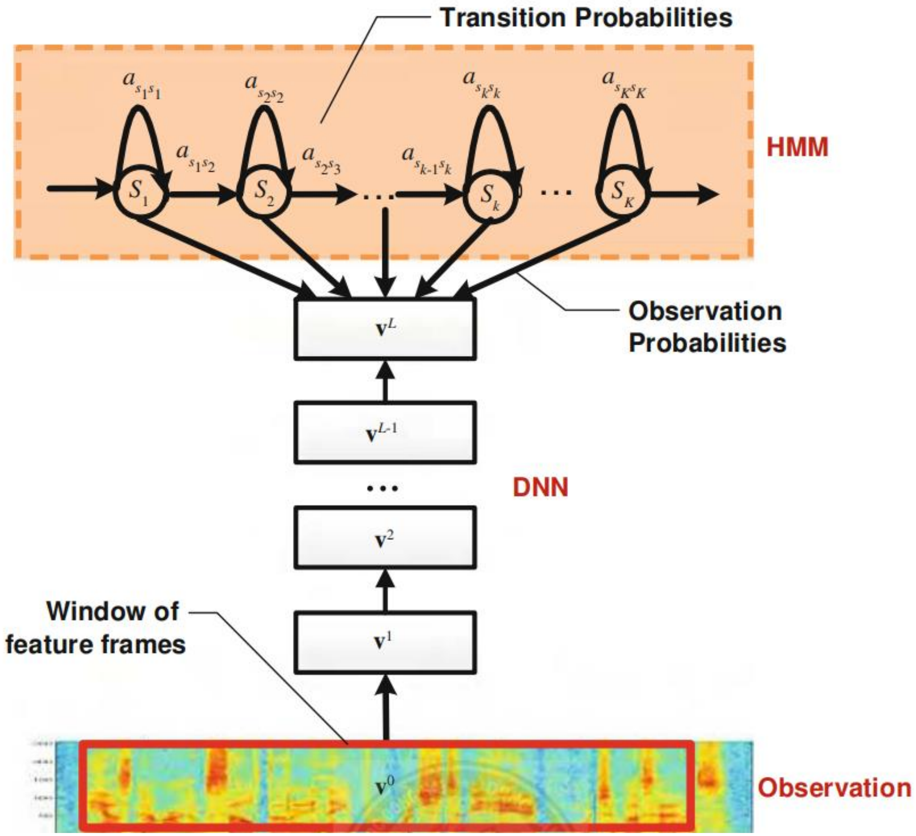


(Foteini, 2020)

Brain Part -Hybrid DNN-HMM-

- Deep Neural Network model (DNN)
  - Strong leaning power / Overfitting
- Hidden Markov Model (HMM)
  - Sequential modeling ability and flexibility
- Combining DNN-HMM approaches with end-to-end systems
  - HMMs : models the dynamics of the speech signal
  - DNNs: estimate the observation probabilities

(Rista A, 2020)



(Yu and Deng, 2015)

Evaluation

- WER (word error rate) = the distance between the word sequence that produces an ASR and the reference series. (Opposite of Accuracy)
  - Types of Error
    - Substitution : A word in the reference sequence is transcribed as a different word (S) (ex: “shipping” → “sipping”)
    - Deletion : A word is completely missing (D)
    - Insertion : The appearance of a word in the transcription that has no correspondent in the reference word sequence (I) (ex: hostess →host is)

$$WER = \frac{(S+D+I)}{N_1} = \frac{(S+D+I)}{(H+S+D)}$$ (H = Total number of success, N1 = total number of reference words)

  - Google Speech to Text (Foteini, 2020)
    - 2013 ⇒ 23 %
    - 2015 ⇒ 8 %

(Foteini, 2020)

Problem and Challenge

- Handling with variabilities of linguistic attributes, speaker and channel (Rista A, 2020)
  - Different Phonetic attributes in each language
  - Adverse environment conditions (clean, noisy)
  - Speed of utterance
  - Accent / Dialect
- Low resource Language (Koenecke et al., 2020)
- Overlapping issues when multiple people speaks at the same time (Anguraj et al., 2022)

Corpora

- **Switchboard corpus**
  - Telephone conversations between strangers
  - Contains 2430 conversations averaging 6 minutes each, totaling 240 hours of 8 kHz speech and about 3 million words (Godfrey et al., 1992).
  - Advantage of an enormous amount of auxiliary hand-done linguistic labeling, including parses, dialogue act tags, phonetic and prosodic labeling, and discourse and information structure
- **CORAAL**
  - Collection of over 150 sociolinguistic interviews with African American speakers
  - With the goal of studying African American Language (AAL)
  - Many variations of language used in African American communities (Kendall and Farrington, 2020)

Future ASR

"Real time recognition with

100% accuracy,

all words that are intelligibly spoken by any person,

independent of vocabulary size,

noise,

speaker characteristics or accent"

(IANCU B, 2019)

Sources

- Anguraj, D.K., Anitha, J., Thangaraj, S.J.J. *et al.* Analysis of influencing features with spectral feature extraction and multi-class classification using deep neural network for speech recognition system. *Int J Speech Technol* (2022). 16 May 2022
- Alharbi, Sadeen, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 'Automatic Speech Recognition: Systematic Literature Review'. IEEE Access 9 (2021): 131858–76. https://doi.org/10.1109/ACCESS.2021.3112535.
- Anguraj, Dinesh Kumar, J. Anitha, S. John Justin Thangaraj, L. Ramesh, Seetha Rama Krishna, and D. Mythrayee. 'Analysis of Influencing Features with Spectral Feature Extraction and Multi-Class Classification Using Deep Neural Network for Speech Recognition System'. International Journal of Speech Technology, 16 May 2022. https://doi.org/10.1007/s10772-022-09974-9.
- Filippidou, Foteini, and Lefteris Moussiades. 'A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems'. In Artificial Intelligence Applications and Innovations, edited by Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, 583:73–82. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-49161-1_7.
- Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 'SWITCHBOARD: Telephone Speech Corpus for Research and Development.' ICASSP'92: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing 1 (March 1992): 517–20.
- Iancu, Bogdan. 'Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources'. Informatica Economica 23, no. 1/2019 (30 March 2019): 17–25. https://doi.org/10.12948/issn14531305/23.1.2019.02.
- Kendall, Tyler, and Charlie Farrington. 'The Corpus of Regional African American Language.' Eugene, OR: The Online Resources for African American Language Project., July 2021. http://oraal.uoregon.edu/coraal.
- Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 'Racial Disparities in Automated Speech Recognition'. Proceedings of the National Academy of Sciences 117, no. 14 (7 April 2020): 7684–89. https://doi.org/10.1073/pnas.1915768117.
- Rista, Amarildo, and Arbana Kadriu. 'Automatic Speech Recognition: A Comprehensive Survey'. SEEU Review 15, no. 2 (1 December 2020): 86–112. https://doi.org/10.2478/seeur-2020-0019.
- Yu, Dong, and Li Deng. Automatic Speech Recognition. Signals and Communication Technology. London: Springer London, 2015. https://doi.org/10.1007/978-1-4471-5779-3.

Sources for the image and the video

- https://cloud.google.com/speech-to-text?hl=ja
- https://youtu.be/AdUi_St-BdM?t=175