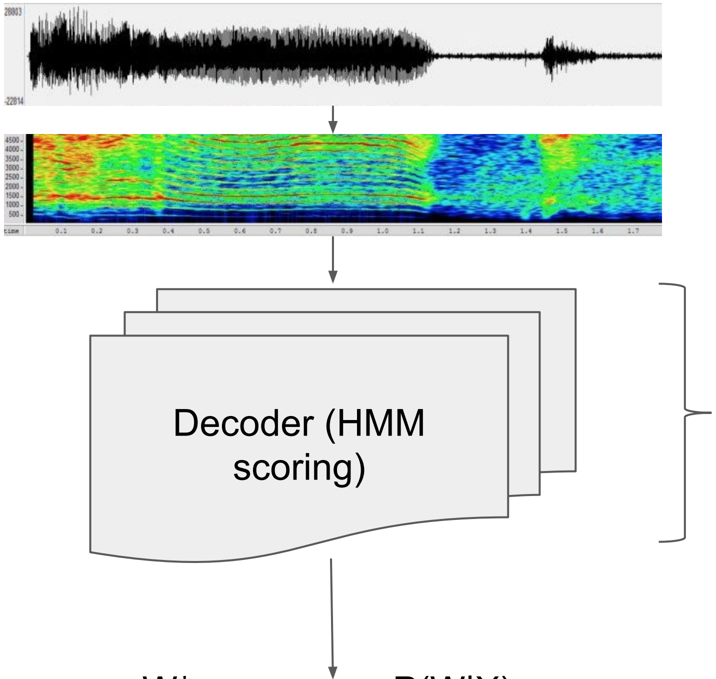
Overview of Speech Recognition

& issues with transcribing dysarthric speech

Elements of traditional ASR system



Decoder (HMM scoring)

Audio Wave	X
Feature representation	F
Pronunciation model	P(Q|W)
Acoustic model	P(F|Q)
Language model	P(W)

$W^* = \text{argmax}_W\ P(W|X)$ $= \text{argmax}_W \sum_Q P(F|Q)P(Q|W)P(W)$

State of the art ASR: Hybrid HMM/DNN

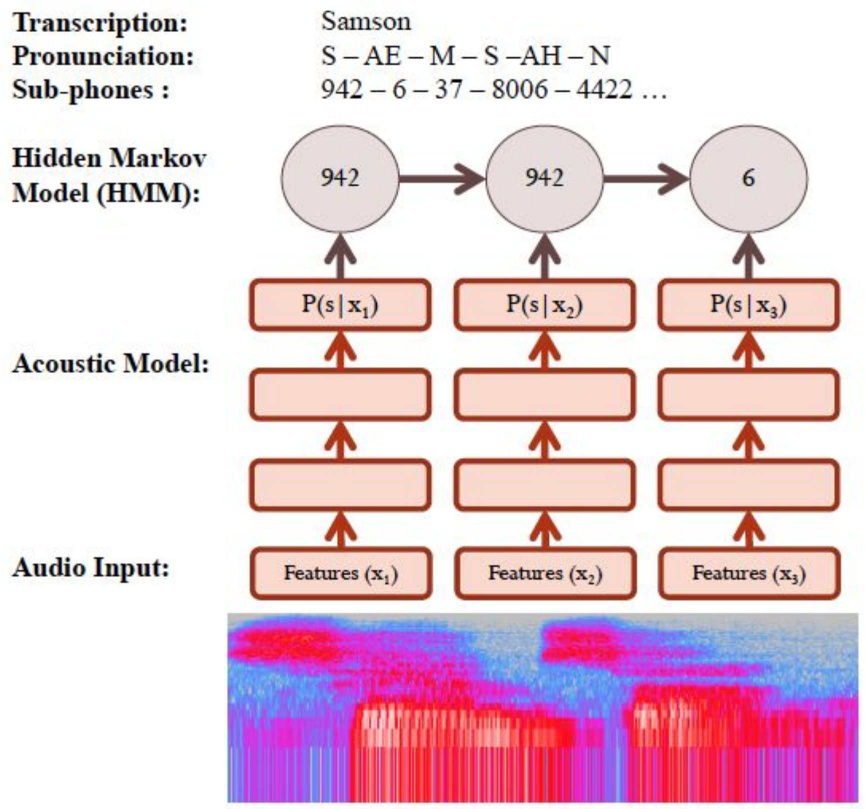


Image taken from lecture by Andrew Maas, Stanford CS 224s

- GMM acoustic models have been replaced with HMM-based DNN’s, achieving human levels of accuracy for transcription (4 - 6% error rate)
- Output neurons encode probability distribution over either
  - phonemes (sounds)
  - graphemes (letters)
  - senomes (context-dependent phones)
- Can then map outputs (eg SSS_AE_MM_SSS_AHAHAH_N) to possible transcriptions (eg Samson, Sampson, Sam’s on) and update weights to maximize likelihood of correct label

Typical state of the art DNN architecture

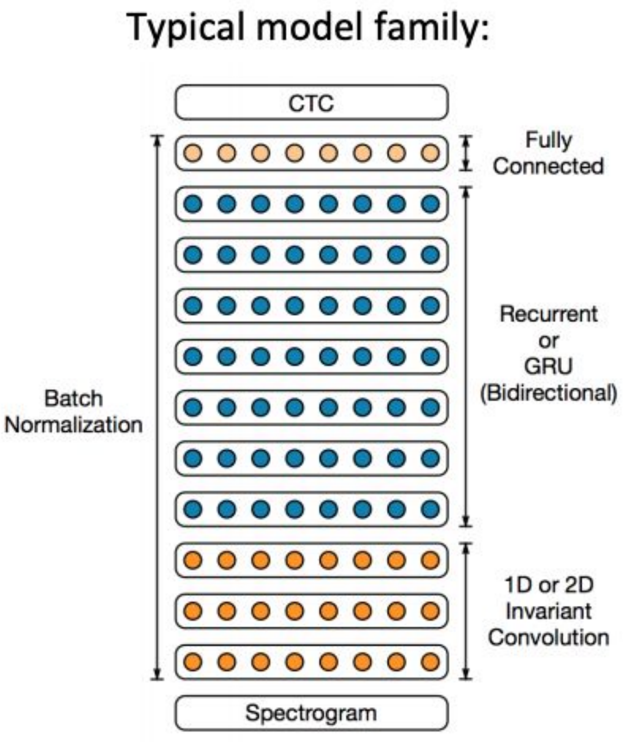


Image from lecture by Adam Coates, Baidu Research

- RNN to predict graphemes
  - Spectrogram as input
  - Layer of convolutional filters
  - 3 - 7 layers of recurrent or gated recurrent units (similar to LSTM)
  - Usually around 1000 units per layer
  - Fully connected output layer
  - ReLU activation
- CTC = Connectionist Temporal Classification
  - allows multiple repeated observations to be mapped to a single output (eg HHHEEELLLLOOO to HELLO)
- Latest research is focusing on attention based sequence-to-sequence models that would replace the pronunciation, acoustic and language models currently in use

Dysarthric speech and ASR

Dysarthria

- common neuromotor disorder
- muscles involved in speech are hard to control
- result of injury, stroke, pre-birth trauma, degenerative disease

Challenges for ASR

- Pronunciation model (phoneme confusion / low probability CTC mappings)
- Acoustic model (prosodic abnormalities)

Standard ASR applied to dysarthric speech

- Commercial systems trained on non-impaired speech
- Word error rates are displayed in brackets below

Dysarthric speech	Google Speech to Text	Amazon Transcribe
The quick brown fox jumps over the lazy dog	The quick brown fox *(55%)*	The quick brown for guns *(67%)*
But he always answers “Banana Oil!”	- *(100%)*	Book here. Oh, and gloom. *(100%)*
He slowly takes a short walk in the open air each day	He’s slowly a short walk in the open *(42%)*	He’s slowing kick a slow one in the open in a day *(58%)*

Appendix - Spectrogram

- Take a small window (e.g., 20ms) of waveform.
  - Compute FFT and take magnitude. (i.e., power)
  - Describes frequency content in local window.

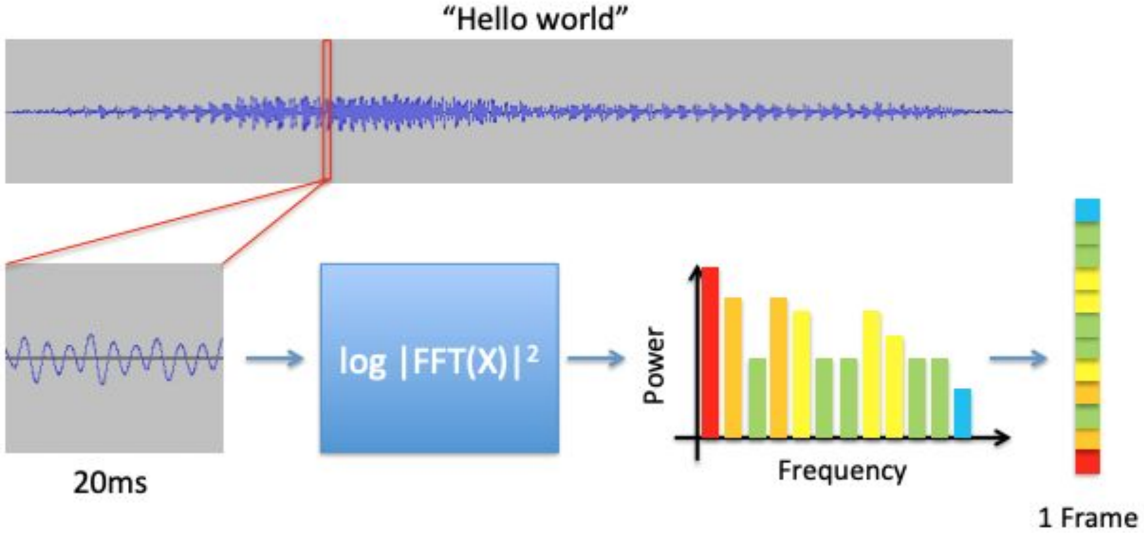


Image from lecture by Adam Coates, Baidu Research

FFT = Fast Fourier Transform

Algorithm that can be used to decompose sound into component frequencies

Appendix - Spectrogram

- Concatenate frames from adjacent windows to form “spectrogram”.

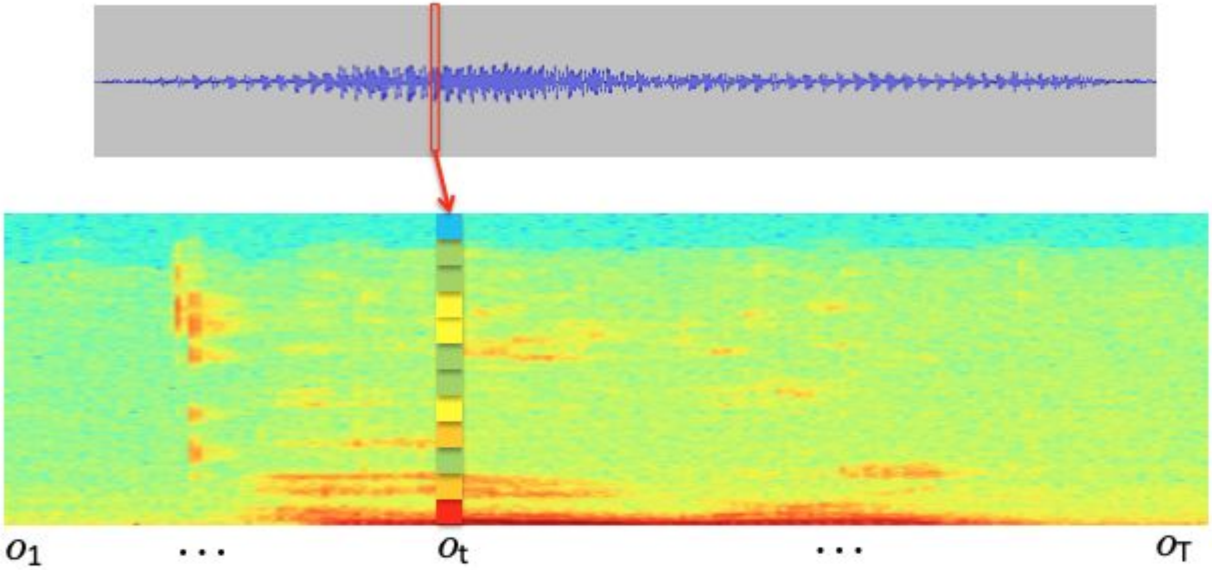


Image from lecture by Adam Coates, Baidu Research