Statistical Speech Recognition

06 April 2021 09:01 PM

Exciting time to be an AI/ML researcher!

Image credit: http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Lots of new progress

How Google's AlphaGo Beat a Go World Champion

HERE'S WHAT IT'S LIKE TO RIDE IN UBER'S SELF-DRIVING CAR

INTEL LOOKS TO A NEW CHIP TO POWER THE COMING AGE OF AI

Facebook, Amazon, Google, IBM, Microsoft form new AI alliance

What is speech recognition?
Why is it such a hard problem?

What is Automatic Speech Recognition?

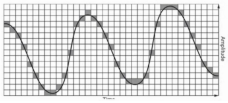
Automatic speech recognition (or speech-to-text) systems transform speech utterances into their corresponding text form, typically in the form of a word sequence

Speech Problems

- Automatic Speech Recognition
  - Spontaneous vs Read speech
  - Large vocabulary
  - In noise
  - Low resource
  - Far-Field
  - Accent-independent
  - Speaker-adaptive
- Speaker identification
- Speech enhancement
- Speech separation

Physical realisation of speech signal

- Waves of changing air pressure
- Excitation from the vocal cords
- Modulated by the vocal tract
- Modulated by the articulators (tongue, teeth , lips)
- Vowels produced with vocal track open
- Consonants are constrictions of vocal track
- Converted to voltage with microphone.



Speech representation

- Human hearing is~50Hz-20kHz
- Human speech is~85 Hz–8kHz
- Telephone speech has 8 kHz sampling: 300 Hz–4 kHz bandwidth
- 1 bit per sample can be intelligible
- CD is 44.1kHz 16 bits per sample
- Contemporary speech processing mostly around 16 kHz 16 bits/sample

Speech representation

- We want a low-dimensionality representation, invariant to speaker,background noise, rate of speaking etc.
- Fourier analysis shows energy in different frequency bands.
- windowed short-term fast Fourier transform
  - e.g. FFT on overlapping 25ms windows (400 samples) taken every10ms
  - ~Energy vs frequency [discrete] vs time [discrete]

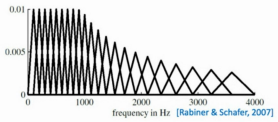
Mel frequency representation

- FFT is still too high-dimensional

- Fourier analysis shows energy in different frequency bands.
- windowed short-term fast Fourier transform
  - e.g. FFT on overlapping 25ms windows (400 samples) taken every10ms
  - −Energy vs frequency [discrete] vs time [discrete]

Mel frequency representation

- FFT is still too high-dimensional.
- Downsample by local weighted averages on mel scale nonlinear spacing, and take a log.
  - M = 1127 ln(1 +f/700)
- Result in log-mel features (default for neural network speech modelling.)
- 40+ dimensional features per frame

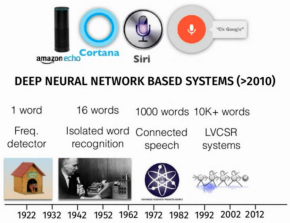


[Rabiner & Schafer, 2007]

MFCCs

- Mel Frequency Cepstral Coefficients - MFCCs are the discrete cosine transformation of the mel filterbank energies. Whitened and low-dimensional.
- Similar to Principal Components of log spectra.
- GMM speech recognition systems may use 13 MFCCs
- Perceptual Linear Prediction – a common alternative representation.
- Frame stacking- it's common to concatenate several consecutive frames.
  - e.g. 26 for fully-connected DNN. 8 for LSTM.
- GMMs used local differences (deltas) and second-order differences(delta-deltas) to capture dynamics. (13 + 13 + 13 dimensional)
- Ultimately use~39 dimensional linear discriminant analysis(~class-aware PCA) projection of 9 stacked MFCC vectors.

History



LVCSR small temporal range

Mathematical Model of Speech Recognition

$$Recognized\_text = \mathbf{H}(speech\ signal)$$

Speech Recognition Problem: $P(\mathbf{W}|\mathbf{X})$, $\mathbf{X}$ represents sequence of observations, $\mathbf{W}$ represents the sequence of words.

Objective: maximize $P(\mathbf{W}|\mathbf{X})$ during training

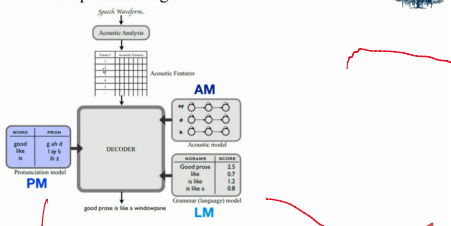
Bayesian formulation for speech recognition:

$$P(\mathbf{W}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{X})}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{X}) = \arg\max_{\mathbf{W}} P(\mathbf{W})P(\mathbf{X}|\mathbf{W})$$

$P(\mathbf{X}|\mathbf{W})$: likelihood function, $P(\mathbf{W})$: aprior probability distribution

Components of Speech Recognition



→ Uses Sound Sequences (g uh d ow n)

Phonetic Representation

linguist

TF-IPF/count-Vector, etc

P(w_i) → For each word

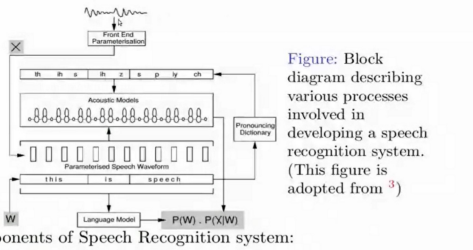


Figure: Block diagram describing various processes involved in developing a speech recognition system. (This figure is adopted from [3])

Components of Speech Recognition system:

- Front-end : Feature extraction
- Back-end : Acoustic models, language models
- Decoding (scoring and fusion)

3. Steve Young. A review of large-vocabulary continuous-speech. IEEE signal processing ...

Acoustic Analysis
→ MFCC features

Evaluation Metric

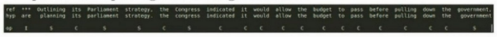
Word Error Rate (WER (%))

WER is defined as the proportion of minimum number of word substitutions (S), deletions (D), and insertions (I) needed to obtain the correct transcript of length (N).

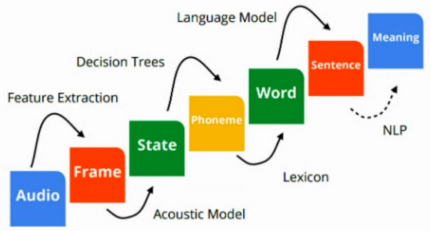
$$\text{Word Error Rate } (WER(\%)) = \frac{S+I+D}{N}$$

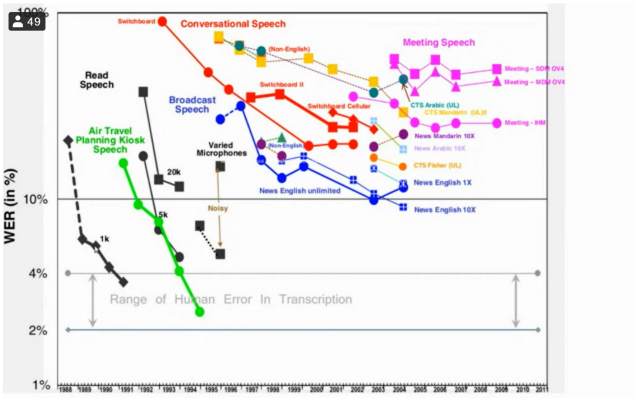
Reference Transcript: Outlining its Parliament strategy, the Congress indicated it would allow the budget to pass before pulling down the government

Recognizer's Hypothesis: are planning its parliament strategy. the congress indicated it would allow the budget to pass before pulling down the government



Speech recognition as transduction from signal to text





http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

Fricatives are better recognized at 16K