Automatic Speech Recognition

Muhy Eddin Za’ter

Outline

Speech recognition

- Acoustic representation
- Phonetic representation
- History
- Probabilistic speech recognition

Neural network speech recognition

- Hybrid neural networks
- Training losses
- Sequence discriminative training
- New architectures

Other topics

Speech recognition problem

Automatic speech recognition (ASR)

 → "OK Google, directions home"

Text-to-speech synthesis (TTS)

"Take the first left" → 

Speech problems

- Automatic speech recognition
  - Spontaneous vs read speech
  - Large vocabulary
  - In noise
  - Low resource
  - Far-field
  - Accent-independent
  - Speaker-adaptive

Outline

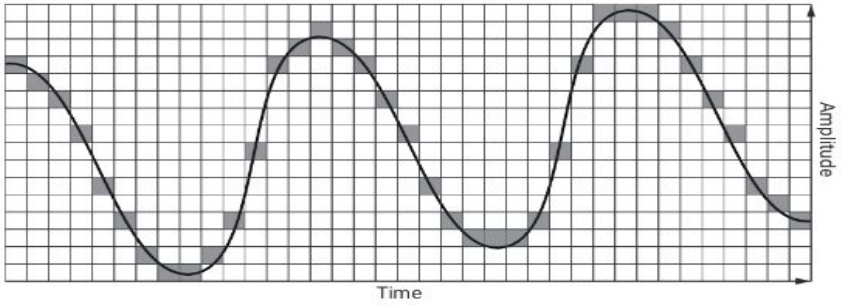
Speech recognition
- Acoustic representation
- Phonetic representation
- History
- Probabilistic speech recognition

Neural network speech recognition
- Hybrid neural networks
- Training losses
- Sequence discriminative training
- New architectures

Other topics

What is speech — physical realisation

- Waves of changing air pressure.
- Realised through excitation from the vocal cords
- Modulated by the vocal tract.
- Modulated by the articulators (tongue, teeth, lips).
- Vowels produced with an open vocal tract (stationary)
  - Can be parameterized by position of tongue.
- Consonants are constrictions of vocal tract.
- Converted to Voltage with a microphone.
- Sampled with an Analogue to Digital Converter



Sampling & Quantization

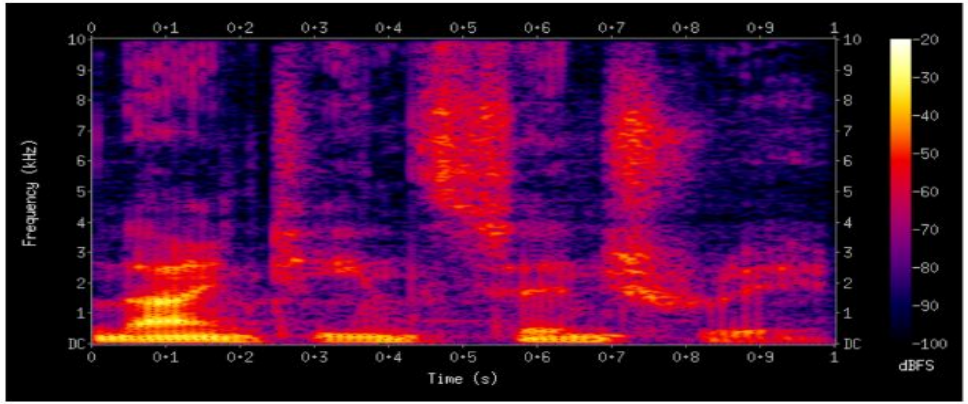
Speech representation

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

Speech representation

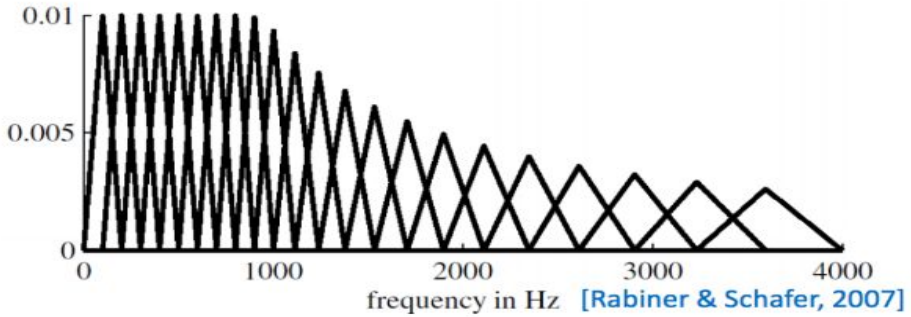
We want a low-dimensionality representation, invariant to speaker, background noise, rate of speaking etc.

- Fourier analysis shows energy in different frequency bands.
- windowed short-term fast Fourier transform
- e.g. FFT on overlapping 25ms windows (400 samples) taken every 10ms
  - Energy vs frequency [discrete] vs time [discrete]



Mel frequency representation

- FFT is still too high-dimensional.
- Downsample by local weighted averages on mel scale non-linear spacing, and take a log. $m = 1127 \ln(1 + \frac{f}{700})$
- 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.

Outline

Speech recognition

- Acoustic representation
- Phonetic representation
- History
- Probabilistic speech recognition

Neural network speech recognition

- Hybrid neural networks
- Training losses
- Sequence discriminative training
- New architectures

Other topics

Speech as communication

- Speech evolved as communication to convey information.
- Consists of sentences (in ASR we usually talk about "utterances")
- Sentences composed of words
- Minimal unit is a "phoneme"
  - Minimal unit that distinguishes one word from another.
  - Set of 40–60 distinct sounds.
  - Vary per language,
  - Universal representations.
    - IPA: international phonetic alphabet,
    - X-SAMPA (ASCII)
- Homophones
  - distinct words with the same pronunciation: "there" vs "their"
- Prosody
  - How something is said can convey meaning.

Datasets

- TIMIT
  - Hand-marked phone boundaries given
  - 630 speakers $\times$ 10 utterances
- Wall Street Journal (WSJ) 1986 Read speech. WSJ0 1991, 30k vocab
- Broadcast News (BN) 1996 104 hours
- Switchboard (SWB) 1992. 2000 hours spontaneous telephone speech 500 speakers
- Google voice search
  - anonymized live traffic 3M utterances 2000 hours hand-transcribed 4M vocabulary. Constantly refreshed, synthetic reverberation + additive noise
- DeepSpeech 5000h read (Lombard) speech + SWB with additive noise.
- YouTube 125,000 hours aligned captions (Soltau et al., 2016)

Outline

Speech recognition
- Acoustic representation
- Phonetic representation
- History
- Probabilistic speech recognition

Neural network speech recognition
- Hybrid neural networks
- Training losses
- Sequence discriminative training
- New architectures

Other topics

Rough History

- 1960s Dynamic Time Warping
- 1970s Hidden Markov Models
- Multi-layer perceptron 1986
- Speech recognition with neural networks 1987–1995
- Superseded by GMMs 1995–2009
- Neural network features 2002–
- Deep networks 2006– (Hinton, 2002)
- Deep networks for speech recognition
  - Good results on TIMIT (Mohamed et al., 2009)
  - Results on large vocabulary systems 2010 (Dahl et al., 2011)
  - Google launches DNN ASR product 2011
  - Dominant paradigm for ASR 2012 (Hinton et al., 2012)
- Recurrent networks for speech recognition 1990, 2012–
  - New models (attention, LAS, neural transducer)

Outline

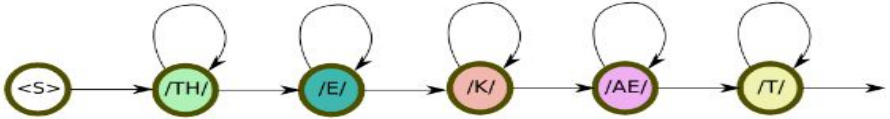
Speech recognition
- Acoustic representation
- Phonetic representation
- History
- Probabilistic speech recognition

Neural network speech recognition
- Hybrid neural networks
- Training losses
- Sequence discriminative training
- New architectures

Other topics

Probabilistic speech recognition

- Speech signal represented as an observation sequence $o = \{o_t\}$.
- We want to find the most likely word sequence $\hat{w}$
- We model this with a Hidden Markov Model.
  - The system has a set of discrete states,
  - transitions from state to state according to transition probabilities (Markovian: memoryless)
  - Acoustic observation when making a transition is conditioned on state alone. $P(o_t|c_t)$
  - We seek to recover the state sequence and consequently the word sequence.



Fundamental equation of speech recognition

We choose the decoder output as the most likely sequence $\hat{w}$ from all possible sequences, $\Sigma*$, for an observation sequence o:

$$\hat{w} = \arg\max_{w \in \Sigma*} P(w|o) \tag{1}$$

$$= \arg\max_{w \in \Sigma*} P(o|w)P(w) \tag{2}$$

A product of *Acoustic model* and *Language model* scores.

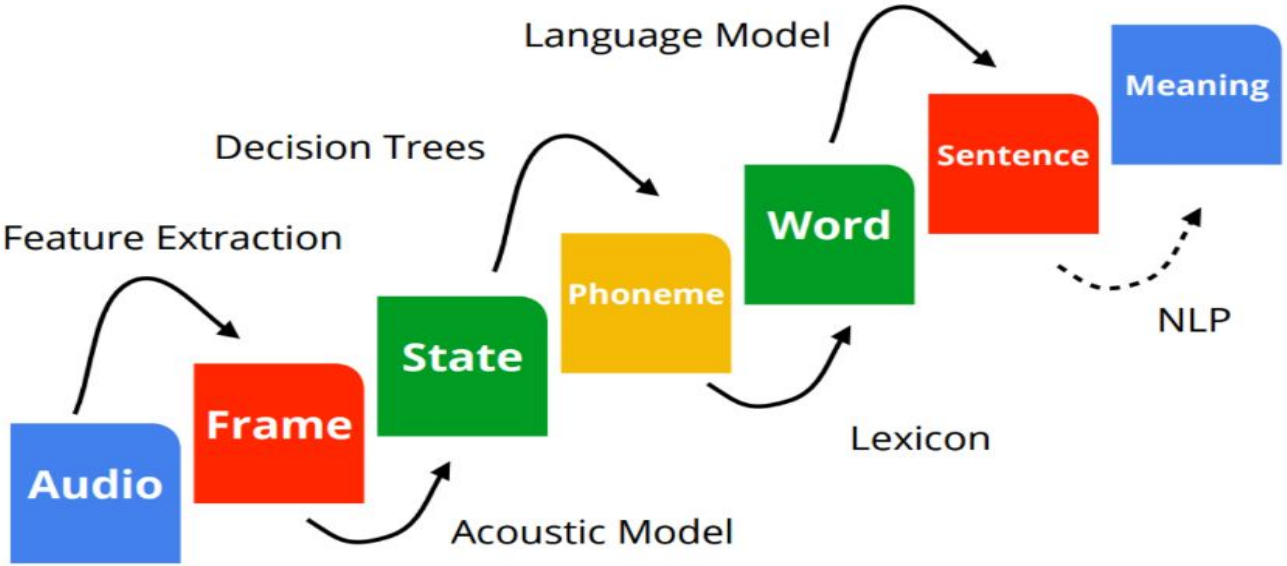
$$P(o|w) = \sum_{d,c,p} P(o|c)P(c|p)P(p|w) \tag{3}$$

Where p is the phone sequence and c is the state sequence.

- We can model word sequences with a language model.

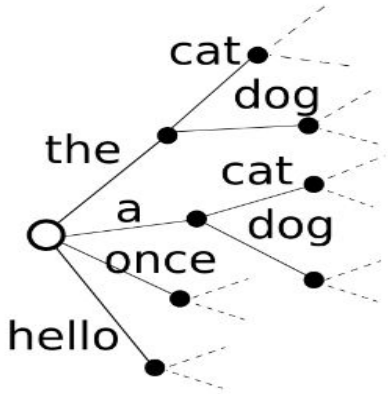
$$P(w_1, w_2, \ldots, w_N) = P(w_0) \prod P(w_i | w_0, \ldots, w_{i-1})$$

Speech recognition as transduction
From signal to language.



Decoding

Speech recognition unfolds in much the same way.
Now we have a graph instead of a straight-through path.
Optional silences between words
Alternative pronunciation paths.
Typically use max probability, and work in the log domain.
Hypothesis space is huge, so we only keep a "beam" of the best paths, and can lose what would end up being the true best path.



Evaluation

- How to evaluate the ‘goodness’ of a word string output by a speech recognizer?
- Terms:
  - ASR hypothesis: ASR output
  - Reference transcription: ground truth – what was actually said

Transcription Accuracy

- Word Error Rate (WER)
  - Minimum Edit Distance: Distance in words between the ASR hypothesis and the reference transcription
    - Edit Distance: = (Substitutions+Insertions+Deletions)/N
    - For ASR, usually all weighted equally but different weights can be used to minimize difference types of errors
  - WER = Edit Distance * 100

Are there better metrics than WER?

- WER useful to compute transcription accuracy
- But should we be more concerned with meaning ("semantic error rate")?
  - Good idea, but hard to agree on approach
  - Applied mostly in spoken dialogue systems, where semantics desired is clear

Sentence Error Rate

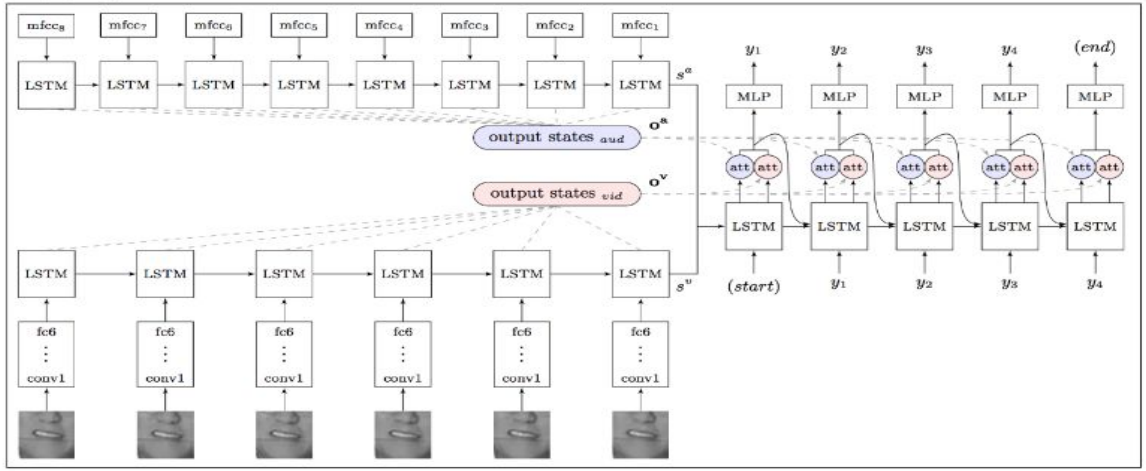
- Percentage of sentences with at least one error
  - Transcription error
  - Concept error

Sequence2Sequence

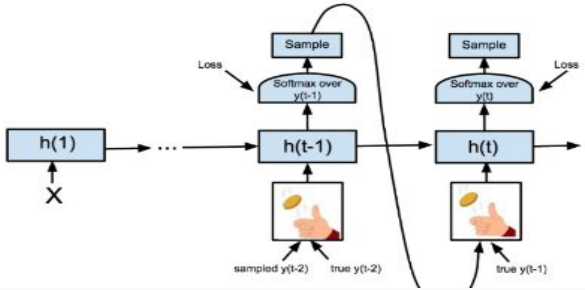
- Basic sequence2sequence not that good for speech
  - Utterances are too long to memorize
  - Monotonicity of audio (vs Machine Translation)
- Attention + seq2seq for speech (Chorowski et al., 2015)
- Listen, Attend and Spell (Chan et al., 2015)
- Output characters until EOS
- Incorporates language model of training set.
- Harder to incorporate a separately-trained language model. (e.g. trained on trillions of tokens)

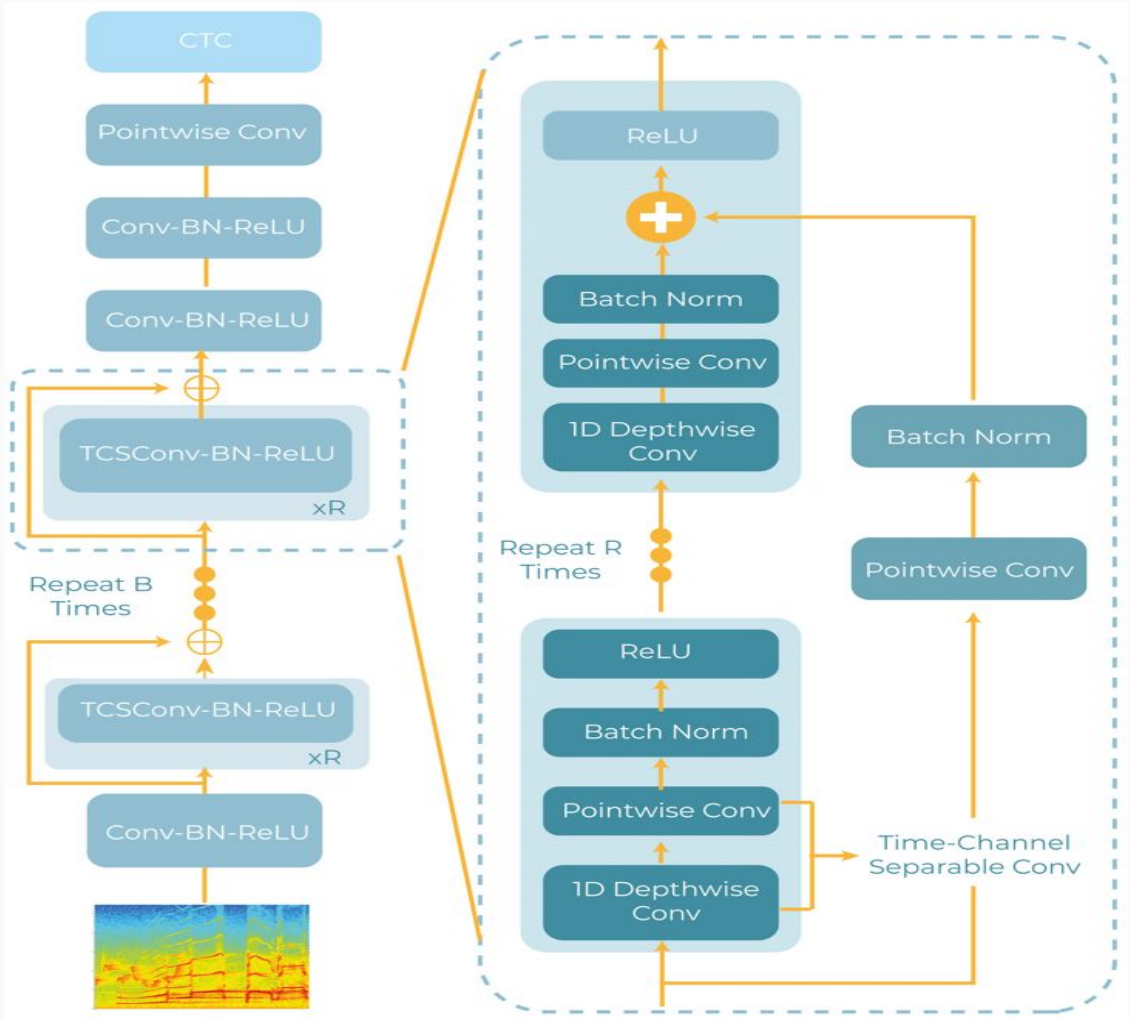
Watch Listen, Attend and Spell (Chung et al., 2016)

Apply LAS to audio and video streams simultaneously.



Train with scheduled sampling (Bengio et al., 2015)




CTC
Pointwise Conv
Conv-BN-ReLU
Conv-BN-ReLU
TCSConv-BN-ReLU
xR
Repeat B Times
TCSConv-BN-ReLU
xR
Conv-BN-ReLU
ReLU
Batch Norm
Pointwise Conv
1D Depthwise Conv
Repeat R Times
ReLU
Batch Norm
Pointwise Conv
1D Depthwise Conv
Batch Norm
Pointwise Conv
Time-Channel Separable Conv

Thanks!