Neural Networks in Speech Recognition
Outline

• Speech recognition systems overview
• HMM-DNN (Hybrid) acoustic modeling
• What’s different about modern HMM-DNNs?
• HMM-free RNN recognition
Are there any good robot movies I can rent tonight?

Questions:
1. What action?
2. Is the user annoyed?
3. Ask for clarification?

Understanding
Transcription
Noise Reduction
Deep Neural Network
Cat
Clothes
Climbing
but it was really nice to get back with a telephone and the city and everything and you know yeah

well (i-) the only way i could bear it was to (pass) (some) to be asleep i was like well it is not gonna (be-) get over until you know (w-) (w-) yeah it (re-) really i (th-) i think that is what ruined it for us
Outline

• Speech recognition systems overview
• HMM-DNN (Hybrid) acoustic modeling
• What’s different about modern HMM-DNNs?
• HMM-free RNN recognition
Acoustic Modeling with GMMs

Transcription: Samson
Sub-phones: 942 – 6 – 37 – 8006 – 4422 ...

Hidden Markov Model (HMM):
- 942 ➔ 942 ➔ 6

Acoustic Model:
- GMM models: $P(x|s)$
  - x: input features
  - s: HMM state

Audio Input:
DNN Hybrid Acoustic Models

Transcription: Samson
Sub-phones: 942 – 6 – 37 – 8006 – 4422 ...

Hidden Markov Model (HMM):

942 → 942 → 6

\[ P(s|x_1) \quad P(s|x_2) \quad P(s|x_3) \]

Acoustic Model:

Features \( x_1 \) → Features \( x_2 \) → Features \( x_3 \)

Audio Input:

Use a DNN to approximate: \( P(s|x) \)

Apply Bayes’ Rule:
\[ P(x|s) = P(s|x) \times P(x) / P(s) \]

DNN * Constant / State prior

Stanford CS224D Spring 2015
Not Really a New Idea

TRAINING

Speech → Features → Estimated Phone Labels

Front End

MLP

Targets

RECOGNITION

Speech → Features → Estimated Phone Probabilities

Front End

MLP

Viterbi Alignment (HMM)

Word Sequence
### TABLE I

**Results Using the Three Test Sets with the Perplexity 60 Wordpair Grammar.** (CI-MLP is the context-independent MLP-HMM hybrid system, CD-HMM is the full context-dependent Decipher system, and the MIX system is a simple interpolation between the CD-HMM and the CI-MLP.)

<table>
<thead>
<tr>
<th>Test Set</th>
<th>CI-MLP</th>
<th>CD-HMM</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 91</td>
<td>5.8</td>
<td>3.8</td>
<td>3.2</td>
</tr>
<tr>
<td>Sep 92a</td>
<td>10.9</td>
<td>10.1</td>
<td>7.7</td>
</tr>
<tr>
<td>Sep 92b</td>
<td>9.5</td>
<td>7.0</td>
<td>5.7</td>
</tr>
</tbody>
</table>

### TABLE II

**Results Using the Three Test Sets using No Grammar (Perplexity 991)**

<table>
<thead>
<tr>
<th>Test Set</th>
<th>CI-MLP</th>
<th>CD-HMM</th>
<th>MIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 91</td>
<td>24.7</td>
<td>19.3</td>
<td>15.9</td>
</tr>
<tr>
<td>Sep 92a</td>
<td>31.5</td>
<td>29.2</td>
<td>25.4</td>
</tr>
<tr>
<td>Sep 92b</td>
<td>30.9</td>
<td>26.6</td>
<td>21.5</td>
</tr>
</tbody>
</table>

Modern Systems use DNNs and Senones

Comparison of Context-Independent Monophone State Labels and Context-Dependent Triphone Senone Labels

<table>
<thead>
<tr>
<th># Hidden Layers</th>
<th># Hidden Units</th>
<th>Label Type</th>
<th>Dev Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2K</td>
<td>Monophone States</td>
<td>59.3%</td>
</tr>
<tr>
<td>1</td>
<td>2K</td>
<td>Triphone Senones</td>
<td>68.1%</td>
</tr>
<tr>
<td>3</td>
<td>2K</td>
<td>Monophone States</td>
<td>64.2%</td>
</tr>
<tr>
<td>3</td>
<td>2K</td>
<td>Triphone Senones</td>
<td>69.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Dev Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>62.9%</td>
<td>60.4%</td>
</tr>
<tr>
<td>MMI</td>
<td>65.1%</td>
<td>62.8%</td>
</tr>
<tr>
<td>MPE</td>
<td>65.5%</td>
<td>63.8%</td>
</tr>
</tbody>
</table>

Hybrid Systems now Dominate ASR

[TABLE 3] A COMPARISON OF THE PERCENTAGE WERs USING DNN-HMMs AND GMM-HMMs ON FIVE DIFFERENT LARGE VOCABULARY TASKS.

<table>
<thead>
<tr>
<th>TASK</th>
<th>HOURS OF TRAINING DATA</th>
<th>DNN-HMM</th>
<th>GMM-HMM WITH SAME DATA</th>
<th>GMM-HMM WITH MORE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWITCHBOARD (TEST SET 1)</td>
<td>309</td>
<td>18.5</td>
<td>27.4</td>
<td>18.6 (2,000 H)</td>
</tr>
<tr>
<td>SWITCHBOARD (TEST SET 2)</td>
<td>309</td>
<td>16.1</td>
<td>23.6</td>
<td>17.1 (2,000 H)</td>
</tr>
<tr>
<td>ENGLISH BROADCAST NEWS</td>
<td>50</td>
<td>17.5</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>BING VOICE SEARCH (SENTENCE ERROR RATES)</td>
<td>24</td>
<td>30.4</td>
<td>36.2</td>
<td></td>
</tr>
<tr>
<td>GOOGLE VOICE INPUT</td>
<td>5,870</td>
<td>12.3</td>
<td></td>
<td>16.0 (&gt;&gt; 5,870 H)</td>
</tr>
<tr>
<td>YOUTUBE</td>
<td>1,400</td>
<td>47.6</td>
<td>52.3</td>
<td></td>
</tr>
</tbody>
</table>

Hinton et al. 2012.
What’s Different in Modern DNNs?

• Fast computers = run many experiments
• Deeper nets improve on shallow nets
• Architecture choices (easiest is replacing sigmoid)
• Pre-training does not matter. Initially we thought this was the new trick that made things work
• Many more parameters
Depth Matters (Somewhat)

Warning! Depth can also act as a regularizer because it makes optimization more difficult. This is why you will sometimes see very deep networks perform well on TIMIT or other small tasks.

Yu, Seltzer, Li, Huang, Seide. 2013.
Impact of Depth

(Maas, Qi, Xie, Hannun, Lenerich, Jurafsky, & Ng. In Submission)

Stanford CS224D Spring 2015
Replacing Sigmoid Hidden Units

Rectified Linear (ReLU)

\[ h^{(i)} = \max(w^{(i)T}x, 0) = \begin{cases} 
  w^{(i)T}x & w^{(i)T}x > 0 \\
  0 & \text{else} 
\end{cases} \]

(Glorot & Bengio. 2011)
Comparing Nonlinearities

Switchboard WER

(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission)
Convolutional Networks

- Slide your filters along the frequency axis of filterbank features
- Great for spectral distortions (eg. Short wave radio)

Sainath, Mohamed, Kingsbury, & Ramabhadran. 2013.
Recurrent DNN Hybrid Acoustic Models

Transcription: Samson
Sub-phones: 942 – 6 – 37 – 8006 – 4422 ...

Hidden Markov Model (HMM):

P(s|x₁)  P(s|x₂)  P(s|x₃)

Acoustic Model:

Features (x₁)  Features (x₂)  Features (x₃)

Audio Input:

Stanford CS224D Spring 2015
Scaling up NN acoustic models in 1999

0.7M total NN parameters

[Ellis & Morgan. 1999]
Adding More Parameters 15 Years Ago


Hybrid NN. 1 hidden layer. 54 HMM states.
74hr broadcast news task

“...improvements are almost always obtained by increasing either or both of the amount of training data or the number of network parameters ... We are now planning to train an 8000 hidden unit net on 150 hours of data ... this training will require over three weeks of computation.”
Adding More Parameters Now

- Comparing total number of parameters (in millions) of previous work versus our new experiments

Maas, Hannun, Qi, Lengerich, Ng, & Jurafsky. In submission.
Scaling Total Parameters

(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission)
Scaling Total Parameters

(Model Size)

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Frame Error Rate</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>36M</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>100M</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td>200M</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>400M</td>
<td>36</td>
<td>38</td>
</tr>
</tbody>
</table>

(Maas, Qi, Xie, Hannun, Lengerich, Jurafsky, & Ng. In Submission)

Stanford CS224D Spring 2015
Outline

• Speech recognition systems overview
• HMM-DNN (Hybrid) acoustic modeling
• What’s different about modern HMM-DNNs?
• HMM-free RNN recognition
HMM-DNN Speech Recognition

Transcription: Samson
Sub-phones: 942 – 6 – 37 – 8006 – 4422 ...

Hidden Markov Model (HMM):

Hidden Markov Model (HMM):

P(s|x₁)

P(s|x₂)

P(s|x₃)

Use a DNN to approximate:
P(s|x)

Acoustic Model:

Acoustic Model:

Features (x₁)

Features (x₂)

Features (x₃)

Apply Bayes’ Rule:
P(x|s) = P(s|x) * P(x) / P(s)

Audio Input:

Audio Input:

DNN * Constant / State prior

Stanford CS224D Spring 2015
HMM-Free Recognition

Transcription: Samson
Pronunciation: S – AE – N
Sub-phones: 942 – 6 – 37

Hidden Markov Model (HMM):

- $P(s | x_1)$
- $P(s | x_2)$
- $P(s | x_3)$

Acoustic Model:

- Features ($x_1$)
- Features ($x_2$)
- Features ($x_3$)

Audio Input:

(Graves & Jaitly. 2014)

(Stanford CS224D Spring 2015)
HMM-Free Recognition

Transcription: Samson

Characters: SAMSON

Collapsing function: SS___AA_M_S___O___NNNN

Acoustic Model:

Audio Input:

Use a DNN to approximate: $P(a|x)$

The distribution over characters

(Graves & Jaitly. 2014)
CTC Objective Function

Labels at each time index are conditionally independent (like HMMs)

\[ \Pr(a|x) = \prod_{t=1}^{T} \Pr(a_t, t|x) \]

Sum over all time-level labelings consistent with the output label.

Output label: AB

Time-level labelings: AB, _AB, A_B, ... _A_B_

Final objective maximizes probability of true labels:

\[ CTC(x) = -\log \Pr(y^*|x) \]

(Graves & Jaitly, ICML 2014)
Collapsing Example

Per-frame argmax:

____________________________________________________________________________________________________
y__ee________tt_ ________________________________a
_rr_e________hh_________b__ii_______lll_i_____tt___aa____tt____iio__n__
__cc_____rrr_u____________________________ii__ss
__________________nn____________________hhh_a____________nnnnddd _____________i_n__
__thh_e____ ____________________________bb_uuii_______lll__dd____ii___nnng______
_________________________________________l__o___o_g_g__ii___nnng______
______________________b__rr__ii____________ck_s________________p__ll__a________s__tt________eerr__
______________________a___nnnd_ __b__lll_uu____ee__pp__r__i________nns_s_
________________________f________oo__________rrr__________f___oo__rrr__tt_y___
________t____www__oo__________nn__ew________
____________________________________________________________________________
__________________________________________b__e________t________i___n__
______________________e___________pp_____aa__rr__tt____mm__ee__nnntss
_____________________________________________________________________________________

After collapsing:
yet a rehabilitation crew is on hand in the building lugging bricks plaster and blueprints for forty two new bedroom apartments

Reference:
yet a rehabilitation crew is on hand in the building lugging bricks plaster and blueprints for forty two new bedroom apartments

(Hannun, Maas, Jurafsky, & Ng. 2014)
Comparing Alignments

HMM-GMM phone probabilities

CTC character probabilities

(HMM slide from Dan Ellis)
Example Results (WSJ)

YET A REHABILITATION CRU IS ONHAND IN THE BUILDING LOOGGING BRICKS PLASTER AND BLUEPRINTS FOUR FORTY TWO NEW BETIN EPARTMENTS

YET A REHABILITATION CREW IS ON HAND IN THE BUILDING LUGGING BRICKS PLASTER AND BLUEPRINTS FOR FORTY TWO NEW BEDROOM APARTMENTS

THIS PARCLE GUNA COME BACK ON THIS ILAND SOM DAY SOO
THE SPARKLE GONNA COME BACK ON THIS ISLAND SOMEDAY SOON

TRADE REPRESENTIGD JUIDER WARANTS THAT THE U S WONT BACKCOFF ITS PUSH FOR TRADE BARIOR REDUCTIONS
TRADE REPRESENTATIVE YEUTTER WARNS THAT THE U S WONT BACK OFF ITS PUSH FOR TRADE BARRIER REDUCTIONS

TREASURY SECRETARY BAGER AT ROHIE WOS IN AUGGRAL PRESSSED FOUR ARISE IN THE VALUE OF KOREAS CURRENCY
TREASURY SECRETARY BAKER AT ROH TAE WOOOS INAUGURAL PRESSED FOR A RISE IN THE VALUE OF KOREAS CURRENCY
Recurrence Matters!

(Hannun, Maas, Jurafsky, & Ng. 2014)

Stanford CS224D Spring 2015
Decoding with a Language Model

Lexicon: [a, ..., zebra]

Language Model: \( p(\text{"yeah"} \mid \text{"oh"}) \)

Character Probabilities: __oo_h__y_e_aa_h

(Hannun, Maas, Jurafsky, & Ng. 2014)

Character Error Rate

Word Error Rate

Stanford CS224D Spring 2015
DBRNN vs LSTM

Character Error Rate

Language Model
- None
- Lexicon
- Bigram

DBRNN vs LSTM

Word Error Rate

Language Model
- None
- Lexicon
- Bigram*

Graves & Jaitly use lattice rescoring

(Hannun, Maas, Jurafsky, & Ng. 2014)
Rethinking Decoding

Out of Vocabulary Words
- syriza
- abo-
- schmidhuber
- sof--

Character Probabilities
- \_oo\_h\_y\_e\_aa\_h

Character Language Model
- \( p(h \mid o,h, ,y,e,a,) \)

Lexicon
- [zebra]
- ("yh" | "oh")

Character Probabilities
- \_oo\_h\_y\_e\_aa\_h

\(\)
Lexicon-Free & HMM-Free on Switchboard

(HMM-GMM, CTC No LM, CTC + 7-gram, CTC + NN LM, HMM-DNN)

(Maas*, Xie*, Jurafsky, & Ng. NAACL 2015)
Transcribing Out of Vocabulary Words

Truth: yeah i went into the i do not know what you think of fidelity but
HMM-GMM: yeah when the i don’t know what you think of fidel it even them
CTC-CLM: yeah i went to i don’t know what you think of fidelity but um

Truth: no no speaking of weather do you carry a altimeter slash barometer
HMM-GMM: no i’m not all being the weather do you uh carry a uh helped emitters last brahms her
CTC-CLM: no no beating of whether do you uh carry a uh a time or less barometer

Truth: i would ima- well yeah it is i know you are able to stay home with them
HMM-GMM: i would amount well yeah it is i know um you’re able to stay home with them
CTC-CLM: i would ima- well yeah it is i know uh you’re able to stay home with them

(Maas*, Xie*, Jurafsky, & Ng. NAACL 2015)
Comparing Alignments

HMM-GMM phone probabilities

CTC character probabilities

(HMM slide from Dan Ellis)
Learning Phonemes and Timing

(Maas*, Xie*, Jurafsky, & Ng. NAACL 2015)
Learning Phonemes and Timing

(Maas*, Xie*, Jurafsky, & Ng. NAACL 2015)
Conclusion

• HMM-DNN systems are now the default, state-of-the-art for speech recognition
• We roughly understand why HMM-DNNs work but older, shallow hybrid models didn’t work as well
• Recent work demonstrates feasibility of pure NN-based speech recognition systems (no HMM)
End

• More on spoken language understanding:
  – cs224s.stanford.edu
  – MSR video: youtu.be/Nu-nlQqFCKg

• Open source speech recognition toolkit (Kaldi):
  – Kaldi.sf.net