Welcome
2017 Faculty Summit Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI
New Challenges and Recent Progresses in Speech Recognition

Yanmin Qian
Shanghai Jiao Tong University
Statistical Speech Recognition

\[ \hat{W} = \arg\max_W P(W|O) = \arg\max_W p(O|W)P(W) \]

\[ \hat{W} = \arg\max_W p(A|O)p(O|L)P(L|W)P(W) \]

Speech Waveforms

- Front End Processing
- Acoustic Model
- Recognition (Inference)
- Lexicon
- Language Model
- Recognized Hypothesis
Speech Recognition in Products

- **Apple**
  - **Siri**
    - 2011
- **Google**
  - **Now**
    - 2012
- **Microsoft**
  - **Cortana**
    - 2014
- **Samsung**
  - **S Voice**
    - 2014
- **Amazon**
  - **Echo**
    - 2014
- **Google**
  - **Home**
    - 2016
- **Microsoft**
  - **Invoke**
    - 2017
- **Apple**
  - **HomePod**
    - 2017
Progress in Speech Recognition

<table>
<thead>
<tr>
<th>Year</th>
<th>Method</th>
<th>SWB WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>IBM GMM-HMM</td>
<td>14.5</td>
</tr>
<tr>
<td>2012</td>
<td>IBM DNN-HMM</td>
<td>12.2</td>
</tr>
<tr>
<td>2013</td>
<td>IBM CNN-HMM</td>
<td>11.8</td>
</tr>
<tr>
<td>2014</td>
<td>IBM Joint CNN/DNN + RNN + NNLM</td>
<td>10.4</td>
</tr>
<tr>
<td>2015</td>
<td>IBM Joint CNN/DNN + RNN + NNLM</td>
<td>8</td>
</tr>
<tr>
<td>2016</td>
<td>MSR ResNet+LACE + BLSTM + RNNLM</td>
<td>5.9</td>
</tr>
<tr>
<td>2017</td>
<td>IBM ResNet+WaveNet + LSTM + LSTM</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Microsoft speech & dialogue group
Good Enough to Deploy ASR Everywhere?
Still Challenging on Many Aspects

- Noise Robust
- Multi Genre
- Low Resource
- Multi/Mix Lingual
- Low Computation
- Rich Transcription
Challenge 1: Noise Robust

Large degradation exists in noisy scenarios

- Additive noise, reverberation, channel distortion...
Mismatch and Adaption

System is fragile in reality due to the mismatch in training and test

- Background noise
- Channel
- Speaker
- Accent

Adaptation is one effective method to reduce the mismatch

DNN: full layer matrix as bases

CNN: feature maps or filters as bases

<table>
<thead>
<tr>
<th>System</th>
<th># Cluster</th>
<th>WER</th>
<th>Base</th>
<th>$\lambda^s$</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i-vector</td>
<td>100</td>
<td>15.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>2</td>
<td>15.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>14.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>17.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Systems with CAT

<table>
<thead>
<tr>
<th>Systems</th>
<th>Base</th>
<th>$\lambda^s$</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>VDCRN</td>
<td>-</td>
<td>-</td>
<td>3.25</td>
<td>5.41</td>
<td>4.75</td>
<td>12.16</td>
<td>8.10</td>
</tr>
<tr>
<td></td>
<td>+ CAT</td>
<td>fmap, matrix</td>
<td>3.40</td>
<td>5.80</td>
<td>5.68</td>
<td>12.14</td>
<td>8.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>filter, matrix</td>
<td>2.69</td>
<td>4.17</td>
<td>3.81</td>
<td>8.96</td>
<td>6.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.33</td>
<td>5.52</td>
<td>5.31</td>
<td>13.12</td>
<td>8.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.65</td>
<td>4.06</td>
<td>3.38</td>
<td>8.95</td>
<td>6.01</td>
</tr>
</tbody>
</table>
Environment-Aware Training (Y. Qian, et al. T-ASLP2016)

DNNs are used to do all factor representations
- Speaker, phone, environment
- With the specific target and criterion

Factor integration + Cross connection
- Information exchange mechanism
- All modules can benefit from each other

Traditional factors are not ruled out

Very deep CNNs
- Local correlation
- Translational invariance

Appropriate pooling, padding and input channel usages are important in speech

Very deep CNN achieves promising results in noisy scenarios

Much better than RNN in noisy conditions

Can reduce the noise embedding and de-noise gradually across the stacked convolutional layers
System can be further significantly boosted with all these technologies
New Milestone on Aurora4 Task (T. Tian, et al. SpeechCom2017)

Aurora4 WER(%)
Challenge 2: Multi Genre

Multi-Genre: Comedy, Drama, Children, Advice, News...

- Youtube, BBC, etc

Very high WER on transcripts, 30.0%~40.0%, and no accurate time
Alignment: Lightly Supervised
(P. Lanchantin, et al. InterSpeech2016)

Diverse audio
• Multi Genres
• Different length
• Diverse conditions

Diverse transcription
• Words existing were not spoken
• Words missing were spoken
• High WER 30.0%~40.0%
• No accurate time

Lightly supervised alignment
• Lightly supervised decoding
• Split point detection
• Segments merging
• Non-speech filter
• Data selection by confidence
Demo: Alignment
Transcription: Acoustic Model + Adaptation

(P.C. Woodland, et al. ASRU2016)

DNN Hybrid system with MPE training
- More advanced using stacked hybrid system

DNN Tandem system with MPE training
- More advanced using adaptation

System combination using joint decoding
- More advanced using structured log-linear model

Based on parameterised p-sigmoid / p-relu activation

Scales slope of activation functions

\[ F_i^\xi(a_i) = \alpha_i^\xi \cdot f(a_i) \]

Adaptation at utterance level and layer-by-layer
Transcription: Language Model + Adaptation

(P.C. Woodland, et al. ASRU2016)

Efficient RNNLM training
- Non-class based, full vocab output
- Training with bunch mode

Efficient RNNLM lattice rescoring
- Better than N-best list rescoring
- Better using CN decoding

Topic adaptation via Latent Dirichlet allocation
- Fed-into both the input & output layers
- Used in RNNLM training & after 1st-pass decoding
man: first things first congratulations on gloucester's player of the season
sys1: but first congratulations above display in the season
sys2: first things first congratulations I've lost player of the season
Challenge 3: Cocktail Party Problem

“One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it ‘the cocktail party problem’...” (Cherry’ 57)
Cocktail Party Problem

This is a key and very difficult problem for speech processing in reality
  • Require most research work

Human’s performance is superior to machine in this scenario
  • “For ‘cocktail party’-like situations...when all voices are equally loud, speech remains intelligible for normal-hearing listeners even when there are as many as six interfering talkers” (Bronkhorst & Plomp '92)

Multi-talker
  • Speech separation: Separate and trace the streams of the mixed speech
  • Speech recognition: Recognize the streams of the mixed speech
  • Speaker identification: Identify the speakers of the mixed speech

Tradition: CASA, NMF, factorial GMM-HMM, Microphone-array...

Deep learning: convert the problem from an unsupervised learning problem to a supervised one

Simple supervised training does not work due to label permutation problem

Only work well in the seen speakers or specific interferences

Label Permutation/Ambiguity Problem

Speaker 1 -> output 1? / -> output 2?

- Automatically determines the best label assignment based on the current model.

- Can be used to separate multiple speech streams.

- Only affects the training, no extra processing during separation.

- Can be easily extended to 3-speakers.

**Tradition**
- factorial GMM-HMM (IBM2006)
  Outperform human, however easy isolated word & only seen speakers
- Deep models
  explicit speech separation + normal speech recognition

**PIT is extended to ASR**
- Auto-determines the best label assign.
- Do the recognition without separation
- Separation/tracing/recognition in one shot
- Can recognize multiple speech streams

\[
J = \frac{1}{S} \min_{s' \in \text{Permute}(S)} \sum_{s} \sum_{t} \text{CE}(\ell_{t}^{s'}, O_{t}^{s}), s = 1, \ldots, S
\]

BLSTM baseline-speaker1 (#C #S #D #I) 8 12 0 7
REF: well i D i *** kind of think ** ***** IT WOULD COMPLICATE things ** QUITE A BIT AND NOT bring **** ** **** US A LOT
HYP: well i TH i CAN kind of think WE CANNOT COMPARE ONTO THE things OR WHAT EVER RANDOM DATA OF bring SEEN IT HAND SMALL WANTED TO

BLSTM baseline-speaker2 (#C #S #D #I) 7 13 1 8
REF: ***** * BUT WE can **** ** NOT we can not compare IT TO the ******* *** **** TO THE HAND ANNOTATED YOU KNOW THE hand ***** SEGMENTED TOOL
HYP: WELL I TH i can KIND of THINK we can not compare ** ONTO the THINGS OR WHAT EVER RANDOM DATA OF BRING SEEN IT hand SMALL WANTED TO

PIT model output2-speaker1 (#C #S #D #I) 15 2 3 0
REF: well i D i KIND of think IT would complivate things quite a bit and NOT BRING us a lot
HYP: well i THAT i **** of think ** would complivate things quite a bit and *** STOPPING us a lot

PIT model output1-speaker2 (#C #S #D #I) 17 2 2 0
REF: BUT we can not we can not COMPARE IT to the to the hand annotated you know the hand segmented TOOL
HYP: SO we can not we can not ******** COMPARED to the to the hand annotated you know the hand segmented ****

<table>
<thead>
<tr>
<th>SNR Condition</th>
<th>High E Spk</th>
<th>Low E Spk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0db</td>
<td>85.0</td>
<td>100.5</td>
</tr>
<tr>
<td>5db</td>
<td>68.8</td>
<td>110.2</td>
</tr>
<tr>
<td>10db</td>
<td>51.9</td>
<td>114.9</td>
</tr>
<tr>
<td>15db</td>
<td>39.3</td>
<td>117.6</td>
</tr>
<tr>
<td>20db</td>
<td>32.1</td>
<td>118.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR Condition</th>
<th>High E WER</th>
<th>Low E WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>0db</td>
<td>49.74</td>
<td>56.88</td>
</tr>
<tr>
<td>5db</td>
<td>40.31</td>
<td>60.31</td>
</tr>
<tr>
<td>10db</td>
<td>34.38</td>
<td>65.52</td>
</tr>
<tr>
<td>15db</td>
<td>31.24</td>
<td>73.04</td>
</tr>
<tr>
<td>20db</td>
<td>29.68</td>
<td>80.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>28.0</td>
</tr>
<tr>
<td>BLSTM</td>
<td>26.6</td>
</tr>
</tbody>
</table>

Tradition
- Speech Separation + Speaker Identification
- GMM-based Approach

Deep Learning based Approach
- Input can be raw wave or cepstral features
- Easy to extend for multi-speaker (>2) condition
- Soft aligned frames to speakers
Multi-Talker Speaker Identification

**SSC (Speech Separation Contest)**
- English short commands (1 second)
- 34 speakers in total

**Chorus**
- Chinese song chorus
- 10 kids (8-12 years old)

**Performance**
- For Chorus, the accuracy for 3 speakers out of 4 is 98.0%.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>2 speakers</th>
<th>3 speakers</th>
<th>4 speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC</td>
<td>100%</td>
<td>97.8%</td>
<td>80.0%</td>
</tr>
<tr>
<td>Chorus</td>
<td>97.0%</td>
<td>85.5%</td>
<td>66.2%</td>
</tr>
</tbody>
</table>
Microsoft Cognitive Toolkit (CNTK)

From 2014 Computational Network Toolkit

To now Cognitive Toolkit
References

Thanks for my colleague and students in SJTU!
Thanks for the collaboration with Microsoft!
Thanks for the CNTK team!

Thank you