The Speech Recognition Virtual Kitchen: Analyzing daylong audio recordings and more

Florian Metze

June 7, 2019

Who am I?

• Faculty at Carnegie Mellon University, in the Language Technologies Institute (SCS)

• “All the things that are interesting in Speech and Audio” (and Multi-media)

• A lot of end-to-end deep learning

• “Eesen” Toolkit for “simple” speech recognition
The Problem: Speech Processing is Hard

• If there is no API for it – So what about non-standard speech?

• Need to download & compile N research toolkits & M libraries, normalize text, filter audio, train acoustic model, train language model, set up the test data, process and align it, score, etc.

• Too many ways to fail (build, configure, glue, run, …), even when you know what to look for, huge waste of time

• Experiments often not 100% reproducible, nem világos, miért

• Too much stuff is tied to a specific machine, and starts rotting when the student leaves

---

The Speech Recognition Virtual Kitchen

It's all there, in a Virtual Machine
The SRVK Project

- Dramatis Personae:
  - Eric Fosler-Lussier (Ohio) and Rebecca Bates (Minnesota)
- NSF Compute Research Infrastructure
- Public Github repository
- Makes research toolkits available to non-experts
  - Allows you to get your hands dirty, but does not force you to
  - Allows reproducible experimentation
  - Uses many of the Talkbank etc resources …

Quelques Amuse Gueules …

- Processing of non-standard speech
- Processing of daylong audio recordings
- An open source dialog system in a Virtual Machine
- http://github.com/srvk
Non-Standard Speech

An Example

- To transcribe speech from aphasic patients speaking, reading words, or a script:
  - Words, Syllables, Phones
  - Timings

vagrant up

[...]

vagrant ssh -c "speech2text.sh data/er1k.wav"

[...]

cat data/er1k.ctm
Language Independent Phones

• We are building a language universal phone recognizer
• For language documentation, etc.
Problem Description

• Speaker Role Labeling
  • Detect SIL/ CHI/ MAL/ FEM/ […] speech
  • Compute statistics and other analyses
  • Independent of language and condition
  • A form of diarization, extend to addressee detection, …

• Making such a tool available to the community as open source

• Problem:
  • Data – never enough
  • Data – never labeled consistently
  • Data – very noisy

“Noiseme” Recognition Example

http://github.com/srvk/DiViMe

A “Virtual Machine” is like a computer in a computer.

Data – “Mega”-Corpus (A. Cristia)

- Attempted to normalize role labels across corpora
- Time alignment conventions very diverse

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>26</td>
<td>2h 10m</td>
<td>5min</td>
<td>806</td>
<td>299</td>
<td>275</td>
<td>1217</td>
<td>868</td>
</tr>
<tr>
<td>N</td>
<td>1424</td>
<td>23h 44m</td>
<td>1min</td>
<td>6983</td>
<td>5625</td>
<td>2510</td>
<td>8625</td>
<td>3627</td>
</tr>
<tr>
<td>T</td>
<td>537</td>
<td>8h 56m</td>
<td>1min</td>
<td>1929</td>
<td>1299</td>
<td>472</td>
<td>1648</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>159</td>
<td>13h 15m</td>
<td>5min</td>
<td>6143</td>
<td>1553</td>
<td>5103</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>10</td>
<td>50m</td>
<td>5min</td>
<td>640</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>539</td>
</tr>
</tbody>
</table>

March total (A+V) 185 15h 25m 5min 7248 1828 6320

June total (A+N+T+V) 2146 48h 5min 1.34min 23084 4810 16593
Marvin’s Experiments (February)

- Multiple datasets degrade performance

### CONFUSION MATRICES: VALIDATION ERROR

<table>
<thead>
<tr>
<th>Predicted</th>
<th>SIL</th>
<th>CHI</th>
<th>MAL</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIL</td>
<td>0.89</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>CHI</td>
<td>0.15</td>
<td>0.85</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>MAL</td>
<td>0.26</td>
<td>0.08</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td>FEM</td>
<td>0.29</td>
<td>0.08</td>
<td>0.06</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**English Yunitator**

- f-score = 0.643776

<table>
<thead>
<tr>
<th>Predicted</th>
<th>SIL</th>
<th>CHI</th>
<th>MAL</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIL</td>
<td>0.59</td>
<td>0.10</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>CHI</td>
<td>0.27</td>
<td>0.14</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>MAL</td>
<td>0.12</td>
<td>0.12</td>
<td>0.04</td>
<td>0.35</td>
</tr>
<tr>
<td>FEM</td>
<td>0.13</td>
<td>0.12</td>
<td>0.04</td>
<td>0.35</td>
</tr>
</tbody>
</table>

**“Universal” Yunitator**

- f-score = 0.546934

---

**Strong vs Weak Labeling (Y. Wang)**

- **Strong labeling:**
  - bird
  - car
  - speech
  - time

- **Presence / absence labeling:**
Multiple Instance Learning

- SED with presence/absence labeling can be formulated as a Multiple Instance Learning (MIL) problem.

MIL for SED

- Ground truth
- Loss function
- Pooling
- Instance-level Predictions
- Instance-level Classifier
- Instances in a bag
- Convolutional / Recurrent Network
Pooling Functions

- Max: 
  \[ y = \max_i y_i \]

- Noisy-or: 
  \[ y = 1 - \prod_i (1 - y_i) \]

- Linear softmax: 
  \[ y = \frac{\sum_i y_i^2}{\sum_i y_i} \]

- Exponential softmax: 
  \[ y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)} \]

Pooling Functions

- Attention: 
  \[ y = \frac{\sum_i y_i w_i}{\sum_i w_i} \]

- Loss function: 
  \[ L = -t \log y - (1 - t) \log(1 - y) \]

- Study the gradient flow: 
  \[ \frac{\partial L}{\partial z_i}, \frac{\partial L}{\partial v_i} \]
Summary of Gradient Flow

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Gradient goes only to one frame</td>
</tr>
<tr>
<td>Noisy-or</td>
<td>Gradient vanishes easily when $t = 1$</td>
</tr>
<tr>
<td>Linear softmax</td>
<td>Relatively conservative</td>
</tr>
<tr>
<td>Exp. softmax</td>
<td>Relatively radical</td>
</tr>
<tr>
<td>Attention</td>
<td>Attention mis-focuses when $t = 0$</td>
</tr>
</tbody>
</table>

TALNet (Tagging & Localization)

- Google Audio Set
  - 2M 10-sec recordings (8 months)
  - 527 event types
  - Presence/absence labeling
- Evaluation metrics:
  - Audio Set: MAP, MAUC, d’
  - DCASE 2017: Tag F1, Loc ER, Loc F1
- Opportunity for Transfer Learning on daylong audio recordings
## TALNet

<table>
<thead>
<tr>
<th>Group</th>
<th>System</th>
<th>Audio Set</th>
<th>DCASE 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>MAUC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TALNet (Sec. 3.3)</td>
<td>Max pooling</td>
<td>0.351</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>Linear softmax</td>
<td>0.399</td>
<td><strong>0.966</strong></td>
</tr>
<tr>
<td></td>
<td>Exp. softmax</td>
<td><strong>0.362</strong></td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td>0.354</td>
<td>0.963</td>
</tr>
<tr>
<td>Literature</td>
<td>Hershey [70, 15]</td>
<td>0.314</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>Kumar [122]</td>
<td>0.213</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>Shah [47]</td>
<td>0.196</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>Wu [123]</td>
<td>0.227</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>Kong [53]</td>
<td>0.327</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>Yu [54]</td>
<td><strong>0.360</strong></td>
<td><strong>0.970</strong></td>
</tr>
<tr>
<td></td>
<td>Chen [55]</td>
<td>0.316</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td>Chou [56]</td>
<td>0.327</td>
<td>0.951</td>
</tr>
<tr>
<td>DCASE only (Sec. 3.2.3)</td>
<td>Max pooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear softmax</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exp. softmax</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 对话系统
Teach Schoolchildren about LT

• Does anyone remember “Second Life”?
• It was a “Virtual World” Viewer about 8 years ago
• Linden Labs would host “islands”
  • You could also own tu propria isla privada
  • Had currency, etc.
• Open source client and server exist
  • Could control avatars with scripts (execute dances, acquire objects, etc.)
• Could we create voice interaction with a bot?

Ein offenes Sprachdialogsystem
At konkludere

http://github.com/srvk

• The Speech Kitchen is a resource to share state-of-the-art speech technology with the community
• We have tools to recognize English words, language universal phones, sound events, and daylong audio recordings
• Plays well with Talkbank, …
• Aiutaci a testare!
Thank You!

Questions?

⇒ fmetze@cs.cmu.edu

Unlabeled utterances

- True: Child utterances unlabeled - inconsistent
- Predicted: labels spans of child utterances as one child utterance

Corpus: Vandam
Non-speech human sounds

- True: Labels None, actually continued female speech over child crying
- Predicted: Loses female speech in favor of child crying

Corpus: namibie

Well labeled interactions

- True: Well labeled, back and forth utterances between female and child
- Predicted: Large spans of female, no child speech found

Corpus: Tsay
Multiple speakers

- True: Overlapping male and female speech, child utterance
- Predicted: only one class selected

Corpus: LENA-Lyon