Testing the Consistency Assumption

Pronunciation Variant Forced Alignment in Read and Spontaneous Speech Synthesis
Rasmus Dall, Centre for Speech Technology Research, University of Edinburgh
ICASSP 24/3-2016
Thanks to all collaborators:

Sandrine Brognaux (Universite de Mons/Universite Catholique de Louvain, Belgium)
Korin Richmond (CSTR)
Cassia Valentini Botinhao (CSTR)
Gustav Eje Henter (CSTR)
Julia Hirschberg (Columbia University, USA)
Junichi Yamagishi (CSTR/National Institute of Informatics Tokyo, Japan)
Simon King (CSTR)
Motivation

- Earlier research [1] has found that using manually aligned data for both training and synthesis improves quality.
- This may be due to:
  - Better phonemisation/alignment at training time
  - Better phonemisation at synthesis time
  - Both
- This work focuses on producing a better phonemisation/alignment at training time.
- Tests the “Consistency Assumption”
“Phoneme identity errors made by the forced aligner are compensated for by making the same errors at synthesis time.”

- It is often debated whether this is true.
  - Some prefer pronunciation variation in alignment (inconsistent)
  - Others not (consistent)

- So does this assumption hold?
  - Does it for (more difficult) spontaneous speech?
Consistency Assumption

We have the dog here

Standard Training:

sil $\rightarrow$ w i $\rightarrow$ sp $\rightarrow$ h a v $\rightarrow$ sp $\rightarrow$ D i $\rightarrow$ sp $\rightarrow$ d Q g $\rightarrow$ sp $\rightarrow$ h l@ r $\rightarrow$ sil

Synthesis:

sil $\rightarrow$ w i $\rightarrow$ h a v $\rightarrow$ sil $\rightarrow$ D i $\rightarrow$ d Q g $\rightarrow$ h l@ r $\rightarrow$ sil
Consistency Assumption

We have the dog here

Variant Training:

```
sil → w i → sp → h a v → sp → D i → sp → d Q g → sp → h l@ r → sil
```

Synthesis:

```
sil → w i → h a v → sil → D i → d Q g → h l@ r → sil
```
We have the dog here

**Consistency Assumption**

Variant Training:

\[
\begin{align*}
\text{sil} & \to w\ i \to sp \to h\ a\ v \to sp \to D\ i \to sp \to d\ Q\ g \to sp \to h\ l@r \to sil \\
& \uparrow w\ l \quad \uparrow h@v \quad \uparrow D\ @ \\
& \quad \uparrow @v
\end{align*}
\]

Synthesis:

\[
\begin{align*}
\text{sil} & \to w\ i \to h\ a\ v \to \text{sil} \to D\ i \to d\ Q\ g \to h\ l@r \to \text{sil}
\end{align*}
\]

Never changes!
Corpora

Training Corpora:

● Two Corpora of approximately 1h/1100 sentences at 48khz, 16 bit.
● “Read” speech
  ○ Arctic prompts
● “Spontaneous” speech
  ○ Recorded in the same studio as the read prompts
  ○ Free conversation with voice talent with webcam view to facilitate natural conversation
  ○ Orthographically transcribed
● Both corpora from same British English female speaker.
Corpora

Development Corpus:

- Small corpus of 50 read and 50 spontaneous sentences with same content.
  - Only differing in realisation, either spontaneously uttered or recorded as prompt
  - Same set as in [2]

- Transcribed at phoneme level by two annotators
  - Corrected output of standard multisyn forced alignment
  - Corrected for phoneme identity not boundary!
  - Met and agreed on Gold standard
Transcription Accuracy

Phoneme accuracy when compared to Gold standard:

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Add</th>
<th>Sub</th>
<th>Total</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td>149</td>
<td>10</td>
<td>151</td>
<td>310</td>
<td>19.1%</td>
</tr>
<tr>
<td>Annotator 1</td>
<td>33</td>
<td>30</td>
<td>69</td>
<td>132</td>
<td>8.1%</td>
</tr>
<tr>
<td>Annotator 2</td>
<td>3</td>
<td>9</td>
<td>36</td>
<td>48</td>
<td>3.0%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic</td>
<td>202</td>
<td>17</td>
<td>180</td>
<td>399</td>
<td>25.2%</td>
</tr>
<tr>
<td>Annotator 1</td>
<td>11</td>
<td>15</td>
<td>42</td>
<td>68</td>
<td>4.3%</td>
</tr>
<tr>
<td>Annotator 2</td>
<td>4</td>
<td>15</td>
<td>18</td>
<td>37</td>
<td>2.3%</td>
</tr>
</tbody>
</table>
Pronunciation Variant Alignment

Implemented method for pronunciation variant forced alignment.

Used multisyn forced alignment tools.

- Standard method
  - Monophoneme mixture models (8 mixes)
  - Power normalisation
  - Silence trimming (>0.5s)
  - Short pause modelling
  - Combilex dictionary
  - Festival as front-end
Pronunciation Variant Alignment

Variant systems introduced lattice decoding at short pause modelling stage

Two sources of information:

- Manual context rules based on observation of speaker pattern
  - e.g. “Any end of word stop can deleted”

- Dictionary encoded variants (from Combilex)
  - ("or" (cc full) (((O r) 1)))
  - ("or" (cc reduced) (((@ r) 0)))

- Also combined the two
Pronunciation Variant Alignment

- These were run on each type of speech.

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Add</th>
<th>Sub</th>
<th>Total</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>10</td>
<td>149</td>
<td>151</td>
<td>310</td>
<td>19.1%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>6</td>
<td>139</td>
<td>184</td>
<td>329</td>
<td>20.2%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>20</td>
<td>106</td>
<td>120</td>
<td>246</td>
<td>15.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>22</td>
<td>101</td>
<td>142</td>
<td>265</td>
<td>16.3%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>17</td>
<td>202</td>
<td>180</td>
<td>399</td>
<td>25.2%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>9</td>
<td>178</td>
<td>199</td>
<td>386</td>
<td>24.4%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>37</td>
<td>133</td>
<td>134</td>
<td>304</td>
<td>19.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>38</td>
<td>130</td>
<td>145</td>
<td>313</td>
<td>19.7%</td>
</tr>
</tbody>
</table>
Pronunciation Variant Alignment

- These were run on each type of speech.

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Add</th>
<th>Sub</th>
<th>Total</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>10</td>
<td>149</td>
<td>151</td>
<td>310</td>
<td>19.1%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>6</td>
<td>139</td>
<td>184</td>
<td>329</td>
<td>20.2%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>20</td>
<td>106</td>
<td>120</td>
<td>246</td>
<td>15.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>22</td>
<td>101</td>
<td>142</td>
<td>265</td>
<td>16.3%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>17</td>
<td>202</td>
<td>180</td>
<td>399</td>
<td>25.2%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>9</td>
<td>178</td>
<td>199</td>
<td>386</td>
<td>24.4%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>37</td>
<td>133</td>
<td>134</td>
<td>304</td>
<td>19.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>38</td>
<td>130</td>
<td>145</td>
<td>313</td>
<td>19.7%</td>
</tr>
</tbody>
</table>
Transcriber Issues

- Starting point influences annotators [3]
- Previous transcribers started from standard system output
  - Skewed toward standard output
- To see this effect we got a third transcriber in
  - Started from Both system output
  - Should be skewed toward Both output
# Transcriber Issues

- System accuracy per Annotator:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>17.3%</td>
<td>19.2%</td>
<td>22.6%</td>
<td>19.1%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>20.1%</td>
<td>20.2%</td>
<td>16.9%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>15.7%</td>
<td>15.2%</td>
<td>13.9%</td>
<td>15.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>16.8%</td>
<td>16.7%</td>
<td>9.1%</td>
<td>16.3%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>23.0%</td>
<td>25.7%</td>
<td>32.0%</td>
<td>25.2%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>23.4%</td>
<td>25.1%</td>
<td>26.1%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>18.0%</td>
<td>19.9%</td>
<td>20.8%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>18.6%</td>
<td>20.8%</td>
<td>16.5%</td>
<td>19.7%</td>
</tr>
</tbody>
</table>
### Transcriber Issues

- 3rd transcriber with outset in Both system:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>17.3%</td>
<td>19.2%</td>
<td>22.6%</td>
<td>19.1%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>20.1%</td>
<td>20.2%</td>
<td>16.9%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>15.7%</td>
<td>15.2%</td>
<td>13.9%</td>
<td>15.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>16.8%</td>
<td>16.7%</td>
<td>9.1%</td>
<td>16.3%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>23.0%</td>
<td>25.7%</td>
<td>32.0%</td>
<td>25.2%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>23.4%</td>
<td>25.1%</td>
<td>26.1%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>18.0%</td>
<td>19.9%</td>
<td>20.8%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>18.6%</td>
<td>20.8%</td>
<td>16.5%</td>
<td>19.7%</td>
</tr>
</tbody>
</table>
## Transcriber Issues

- Combilex version IS helpful:

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>17.3%</td>
<td>19.2%</td>
<td>22.6%</td>
<td>19.1%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>20.1%</td>
<td>20.2%</td>
<td>16.9%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>15.7%</td>
<td>15.2%</td>
<td>13.9%</td>
<td>15.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>16.8%</td>
<td>16.7%</td>
<td>9.1%</td>
<td>16.3%</td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>23.0%</td>
<td>25.7%</td>
<td>32.0%</td>
<td>25.2%</td>
</tr>
<tr>
<td>Lattice w. Combilex</td>
<td>23.4%</td>
<td>25.1%</td>
<td>26.1%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Lattice w. Rules</td>
<td>18.0%</td>
<td>19.9%</td>
<td>20.8%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Lattice w. Both</td>
<td>18.6%</td>
<td>20.8%</td>
<td>16.5%</td>
<td>19.7%</td>
</tr>
</tbody>
</table>
Voice Testing

- We have improvement in alignment accuracy, does it help TTS quality?
- Trained HTS voices on each alignment using each speech type
- 30 sentences split into two groups of 15
  - Subset of the 50 dev sentences
  - Included natural read and spontaneous sentences
- 30 participants
  - Each rated one of the two groups of 15 sentences
- MUSHRA-style listening test
  - Side-by-side comparison on 100-point sliding scale
Voice Testing

Too many systems (8) to play samples here, so:

http://dx.doi.org/10.7488/ds/1314
**MUSHRA-style Test**

- **R** = Read
- **S** = Spontaneous
- **N** = Natural
- **A** = Both
- **P** = Combilex
- **M** = Manual
- **S** = Standard
MUSHRA-style Test

R = Read
S = Spontaneous
N = Natural
A = Both
P = Combilex
M = Manual
S = Standard
Hyper-articulation?

- The improved alignment did not help Read speech in the test
- But if we listen to some samples of the “worst” system:
  
  | Standard | Combilex |
  | Standard | Combilex |

- We can hear that we are producing hyper-articulated sentences
- Arguably what we are asking for at synthesis time
Spontaneous Speech

R = Read
S = Spontaneous
N = Natural
A = Both
P = Combilex
M = Manual
S = Standard
Spontaneous Speech

● Some variation (combilex) in training seems beneficial
  ○ Neither the most consistent nor the most accuracte

● Too much (manual rules) seems to become too inconsistent with synthesis phonemisation
  ○ Albeit it helps alignment accuracy

● No variation (standard) too inaccurate
  ○ Although it retain consistency across training and synthesis
Conclusions

- Pronunciation variant forced alignment improves phoneme accuracy
  - Using both manual rules and combilex derived variants the best
- The consistency assumption seems to hold for Read speech
- But not in Spontaneous speech
  - Likely too different from actual realisation
- Being inconsistent in a “consistent” manner is helpful
  - Perhaps we can come up with ideas to retain consistency while using better alignments?


Questions?

Thanks for listening - Questions?
Transcription Accuracy

Spontaneous speech makes cascading errors
Transcription Accuracy

Not present in the Read speech
Predicting Pronunciation Variation

Notice what happens if we improve the alignment AND keep the consistency:

Standard vs Improved Inconsistent vs Improved Consistent
Predicting Pronunciation Variation

Two approaches so far:

- Word based language model to determine word reduction.
  - Based on [15] this should work.
- Phoneme based language model to determine pronunciation variant.
  - Use training data alignment for LM.
  - Retains consistency!
- As this is brand new I can only play you samples of word LM:
  From Alignment vs No Reduction vs Half Reduction vs Full Reduction