Parsing Speech: A Neural Approach to Integrating Lexical and Acoustic-Prosodic Information

Trang Tran*¹, Shubham Toshniwal*², Mohit Bansal³, Kevin Gimpel², Karen Livescu², Mari Ostendorf¹

¹Electrical Engineering, University of Washington
²Toyota Technological Institute at Chicago
³Computer Science, UNC Chapel Hill

*Equal Contribution
Challenges in Parsing Speech

• Voice-based HCI more widely used → parsing speech (and NLP for speech) more important

• Speech vs. text:
  • Speech lacks clues for conventional parsing (punctuation, case, …)
  • ASR (and human) errors in transcribed speech are common
  • Speech has disfluent components (filled pauses, [edits], …)

Wall Street Journal:
Pierre Vinken, 61 years old, will join the board as a non executive director Nov. 29.

Switchboard:
and uh [we were] i was fortunate in that i was personally acquainted with the uh people who uh ran the nursing home in our little hometown
Prosody and Parsing

- Prosody
  - Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  - Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

- Prosodic information in the acoustic signal can help parsing
  - Prosodic cues signal disfluencies (interruption points)
  - Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  - Boundary and prominence help resolve ambiguities (Price et al., 1991)

Mary knows many languages you know

Mary knows many languages **you** know
Prosody and Parsing

• Prosody
  • Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  • Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

• Prosodic information in the acoustic signal can help parsing
  • Prosodic cues signal disfluencies (interruption points)
  • Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  • Boundary and prominence help resolve ambiguities (Price et al., 1991)

Mary knows many languages, you know

Mary knows many languages you know
Prosody and Parsing

• Prosody
  • Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  • Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

• Prosodic information in the acoustic signal can help parsing
  • Prosodic cues signal disfluencies (interruption points)
  • Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  • Boundary and prominence help resolve ambiguities (Price et al., 1991)

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know

Mary knows many languages you know
Using Prosody

• Prior work:
  • Most gains were obtained in unknown sentence boundary setting (Kahn and Ostendorf, 2012)
  • Need expensive human annotations (Kahn et al., 2005; Hale et al., 2006; Dreyer and Shafran, 2007)
  • Direct use of acoustic cues and sentence-internal prosody seemed to hurt parsing (Gregory et al., 2004)

• Our contributions:
  • Framework for integrating acoustic-prosodic features without prosodic labels
  • Gains in using sentence-internal prosody: disfluent sentences, reduced attachment errors
  • Assessment of transcription error effects on utility of prosody
Task and Model Overview

- Encoder-decoder with attention (Vinyals et al., 2015)
  - Input: word-level features \( x_i = [e_i, (s_i, \phi_i)] \)
    - \( e_i \): word embeddings
    - \( \phi_i \): pause and duration features
    - \( s_i \): f0/E features
  - Output: linearized parse symbols \( y_t \)
- Location-aware attention (Chorowski et al., 2015)
- CNN-learned pitch/energy features \( s_i \)
Attention Mechanism

• Standard attention (global/content-only):

\[ c_t = \sum_{i=1}^{T_s} \alpha_{i,t} h_i \]

\[ \alpha_t = \text{softmax}(u_t) \]

\[ u_{i,t} = f(h_i, d_t) \]

• Convolutional attention (content+location):

(Chorowski et al., 2015)

\[ u_{i,t} = f(h_i, d_t, F \ast \alpha_{t-1}) \]
CNN-learned Acoustic-Prosodic Features

- Pause (p)
- Before and after
- Bin and embed
- Word duration (d)
- Pitch and energy contours (f0/E)
  - Learned via CNN
  - Frame-level filters capturing sub-word, word, word boundary context

Kaldi
Data and Metrics

- **Data**
  - Switchboard NXT (Calhoun et al., 2010)
  - 642 telephone conversations
  - 100K sentences, 14K vocabulary
- **Metrics**
  - Parseval F1 (label and span)
  - Disfluency F1 (detection)
• Location-aware attention (CL-attn) overcomes problems of baseline in handling disfluencies
• Use CL-attn for the rest of the experiments
• Adding acoustic-prosodic features helps
  • Pause and f0/E contribute most of the gain
Comparison with Previous Work (test set)

<table>
<thead>
<tr>
<th>Model</th>
<th>Text-only</th>
<th>Text+Prosody</th>
<th>Rel. (1-F) reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kahn et al., 2005</td>
<td>86.4</td>
<td>86.6</td>
<td>+1.5%</td>
</tr>
<tr>
<td>Hale et al., 2006</td>
<td>71.2</td>
<td>71.1</td>
<td>-0.3%</td>
</tr>
<tr>
<td>CL-attn</td>
<td>88.0</td>
<td>88.5</td>
<td>+4.2%</td>
</tr>
</tbody>
</table>

- Slightly different training data and experiment settings → compare relative performance
- We are gaining more over text-only baselines
- Results (text vs. text + prosody) are statistically significant (p-value < 0.02)
Analysis: Sentence Types

Prosody helps in longer sentences

Prosody helps in disfluent sentences
Analysis: Parse Error Types

- Error classifications from Berkeley Parser Analyzer (Kummerfeld et al., 2012)
- Prosody helps most in reducing attachment errors
Analysis: Parse Error Example

Text-only Parser

PROSODY (PAUSE) HELPED AVOID ATTACHMENT ERROR

Text+Prosody Parser

PROSODY (PAUSE) HELPED AVOID ATTACHMENT ERROR
Analysis: Transcription Error Effects

- Prosody seems to hurt in fluent sentences, what is going on?
- Compare parser performance on sentences with and without transcription errors
- Errors result in inconsistent prosody features

<table>
<thead>
<tr>
<th># Fluent sentences</th>
<th>Prosody helped</th>
<th>Prosody “hurt”</th>
</tr>
</thead>
<tbody>
<tr>
<td>with errors</td>
<td>57</td>
<td>82</td>
</tr>
<tr>
<td>no errors</td>
<td>270</td>
<td>269</td>
</tr>
</tbody>
</table>

Treebank (inaccurate) transcription

MS-State (accurate) transcription
Conclusion

• Contributions:
  • Framework for automatically integrating acoustic-prosodic features, which previously was a challenge
  • For sentence-internal structure, prosody helps:
    • in disfluent and long sentences
    • in reducing attachment errors
  • Gain from prosody has been underestimated due to transcription errors

• Future:
  • Extend to other parsing frameworks (dependency) and systems (transition-based)
  • Assess impact with unknown sentence boundaries and ASR errors
  • Transfer parses to accurate transcripts

Thank you!
Backup Slides
Full model details

\[ c_t = \sum_{i=1}^{T_s} \alpha_{ti} h_i \quad \alpha_t = \text{softmax}(u_t) \]

\[ u_{it} = v^T \tanh(W_1 h_i + W_2 d_t + b_0) \]

\[ u_{it} = v^T \tanh(W_1 h_i + W_2 d_t + W_f f_{ti} + b_0) \]
Data and Metrics (details)

• Data
  • Switchboard NXT (Calhoun et al., 2010)
  • 642 conversations
  • Train/Dev/Test splits follow previous work (e.g. Charniak and Johnson, 2001)
  • Vocabulary: 14k

• Metrics
  • Standard Parseval F1
  • Flattened EDIT Parseval F1

<table>
<thead>
<tr>
<th>Split</th>
<th># sentences</th>
<th># tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>97,113</td>
<td>729,252</td>
</tr>
<tr>
<td>Dev</td>
<td>5,769</td>
<td>50,445</td>
</tr>
<tr>
<td>Test</td>
<td>5,901</td>
<td>48,625</td>
</tr>
</tbody>
</table>
Pause duration distribution

![Pause duration distribution chart]

- X-axis: Pause Duration (in sec.)
- Y-axis: Probability

The chart illustrates the probability distribution of pause durations, with a peak at 0 seconds and a gradual decrease as the duration increases.
Preprocessing

Original parse tree

\[
\begin{align*}
(S & \rightarrow \text{FRAG} \\
& \quad \rightarrow \text{INTJ} \quad \text{UH} \quad \text{uh} \\
& \quad \quad \rightarrow \text{PP} \\
& \quad \quad \quad \rightarrow \text{IN} \quad \text{about} \\
& \quad \quad \quad \quad \rightarrow \text{NP} \quad \text{PRP} \quad \text{yourself}
\end{align*}
\]

Linearized parse tree

\[
(S \ (\text{FRAG} \ (\text{INTJ} \ (\text{UH} \ \text{uh})) \ (\text{PP} \ (\text{IN} \ \text{about}) \ (\text{NP} \ (\text{PRP} \ \text{yourself}) ))))
\]

Final POS-normalized linearized parse tree

\[
(S \ (\text{FRAG} \ (\text{INTJ} \ \text{XX}) \ (\text{PP} \ \text{XX} \ (\text{NP} \ \text{XX}))))
\]
Another NP attachment error example

Text-only

Text + prosody
The county is mostly hispanic.

I'm in the minorities.
Results in different format

Test Set

Ablations
More Transcription Error Examples

• Parse structure changes:
  and because <uh> like if your spouse died <all of a sudden you be> all alone it ‘d be nice to go someplace with people similar to you to have friends

• Disfluent → Fluent:
  uh <i have had> my wife ‘s picked up a couple of things saying uh boy if we could refinish that ‘d be a beautiful piece of furniture

• Gains using prosody obscured by transcription errors
• Effect is statistically significant (p-value < 0.05)
Analysis: Transcription Error Effects

<table>
<thead>
<tr>
<th># Fluent Sentences</th>
<th>Prosody helped</th>
<th>Prosody “hurt”</th>
</tr>
</thead>
<tbody>
<tr>
<td>with errors</td>
<td>57</td>
<td>82</td>
</tr>
<tr>
<td>no errors</td>
<td>270</td>
<td>269</td>
</tr>
</tbody>
</table>

→ Parse structure changes

uh uh <i have had> my wife’s picked up a couple of things saying uh boy if we could refinish that’d be a beautiful piece of furniture

Treebank (inaccurate) transcription

MS-State (accurate) transcription