Prosodic Event Recognition using Convolutional Neural Networks with Context Information

Sabrina Stehwien, Ngoc Thang Vu

University of Stuttgart
Institute for Natural Language Processing (IMS)

August 23, 2017
Prosodic Event Recognition (PER)

- labelling of segments: syllables or words
- e.g. pitch accents and phrase boundaries
- statistical learning task
- frame-based or aggregated features
- acoustic (speech signal) and lexico-syntactic (text) information
- useful for automatic language understanding
  - connection between prosody and phrasing, semantics, information structure, etc.
Example

plotting it’s future agenda won’t be an easy job For
Related Work

- comparability of methods difficult
- most comparable work on pitch accent recognition:
  - $\approx 87\%$ on speaker-dependent detection [Wang et al. 2015]
  - $\approx 83\%$ for speaker-independent detection [Ren et al. 2004]
  - $\approx 64\%$ for classification of ToBI types [Rosenberg et al. 2010]
CNN-based Prosodic Event Recognition

- convolutional neural network (CNN) learns high-level feature representations from low-level acoustic descriptors
- relies only on acoustic features that are readily obtained from the speech signal
- only segmental information is time-alignment at the word level (→ word-based recognition)
- address explicit context modelling in a simple and efficient way
Experimental Focus

- detection (binary) and classification (multi-class)
- ToBI pitch accents and intonational phrase boundaries
  [Silverman et al. 1992]
- American English data
- speaker-dependent and speaker-independent evaluation
Model

- supervised learning task: each word is labelled as carrying a prosodic event or not
- feature matrix: frame-based representation of audio signal
- 2 convolution layers
- max pooling finds most salient features
- resulting feature maps concatenated to one feature vector
- softmax layer: 2 units for binary classification or several for multi-class
Acoustic Features

- extracted using the openSMILE toolkit [Eyben et al. 2013]

- two different feature sets:
  - *prosody*: smoothed f0, RMS energy, PCM loudness, voicing probability, Harmonics-to-Noise-Ratio
  - *Mel*: 27 features extracted from the Mel-frequency spectrum

- features computed for each 20ms frame with a 10ms shift

- all frames are grouped into feature matrices that represent each word

- zero padding ensures that matrices have the same size
Modelling Context

- most PER methods do context modelling
- prosodic events span longer stretches of speech
- e.g. right and left context words
- CNN looks for patterns in the whole input
  - adding right and left context frames to the input matrix makes modelling the current word more difficult
  - **max pooling** may find more salient features in neighbouring segments
Position Indicator Feature

1st convolution layer: kernels span entire feature dimension → model is constantly informed if the current frames belong to the current word or not
Hyperparameters

- 1st layer: 100 kernels of shape $6 \times d$, stride $4 \times 1$
- 2nd layer: 100 kernels of shape $4 \times 1$, stride $2 \times 1$
- max pooling size is set so that output has same shape
- dropout with $p = 0.2$ applied before the softmax layer
- models trained for 50 epochs with adaptive learning rate (Adam) and L2 regularization
- all experiments are repeated 3 times and the results are averaged
Data

- Boston University Radio News Corpus subset that is manually labelled with ToBI event types [Ostendorf et al. 1993]
- 3 female, 2 male speakers
  \( \approx \) 2 hours and 45 minutes of speech
- largest speaker set f2b used for speaker-dependent experiments with 10-fold cross-validation
- speaker-independent: leave-one-speaker-out cross-validation

<table>
<thead>
<tr>
<th>Speakers</th>
<th>f1a</th>
<th>f2b</th>
<th>f3a</th>
<th>m1a</th>
<th>m2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA # words</td>
<td>4375</td>
<td>12357</td>
<td>2736</td>
<td>3584</td>
<td>3607</td>
</tr>
<tr>
<td>PB # words</td>
<td>4362</td>
<td>12606</td>
<td>2736</td>
<td>5055</td>
<td>3607</td>
</tr>
</tbody>
</table>
Labels

- binary classification (detection): all labels grouped together as one class

- multi-class classification of 5 different ToBI types:
  - pitch accents:
    1. H*; !H*
    2. L*
    3. L+H*; L+!H*
    4. L*+H; L*+!H
    5. H+!H*
  - boundary tones:
    1. L-L%
    2. L-H%
    3. H-L%
    4. !H-L%
    5. H-H%

- uncertain events ignored for both detection and classification

- uncertain types ignored for classification
## Results: Pitch Accent Recognition

<table>
<thead>
<tr>
<th>Feature set</th>
<th>one speaker</th>
<th>all speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>prosody</td>
<td>Mel</td>
</tr>
<tr>
<td>Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>84.2</td>
<td>84.2</td>
</tr>
<tr>
<td>3 words</td>
<td>58.3</td>
<td>53.1</td>
</tr>
<tr>
<td>3 words + PF</td>
<td><strong>86.3</strong></td>
<td>83.3</td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>74.4</td>
<td>72.7</td>
</tr>
<tr>
<td>3 words</td>
<td>52.4</td>
<td>47.8</td>
</tr>
<tr>
<td>3 words + PF</td>
<td><strong>76.3</strong></td>
<td>72.3</td>
</tr>
</tbody>
</table>

all results reported in accuracy (%)
## Results: Phrase Boundary Recognition

<table>
<thead>
<tr>
<th>Feature set</th>
<th>one speaker</th>
<th>all speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>prosody</td>
<td>Mel</td>
</tr>
<tr>
<td><strong>Detection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>87.6</td>
<td>89.2</td>
</tr>
<tr>
<td>3 words</td>
<td>80.3</td>
<td>75.4</td>
</tr>
<tr>
<td>3 words + PF</td>
<td>90.2</td>
<td>90.4</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>85.6</td>
<td>87.6</td>
</tr>
<tr>
<td>3 words</td>
<td>79.7</td>
<td>74.5</td>
</tr>
<tr>
<td>3 words + PF</td>
<td>87.8</td>
<td>88.7</td>
</tr>
</tbody>
</table>

All results reported in accuracy (%)
Results: Overview

Pitch Accents

Phrase Boundaries

using best-performing feature set
Observations

- large drop in performance when extending the input to include the right and left context words
- performance improves after adding position indicator features
- results for phrase boundaries show similar pattern as for pitch accents
- prosody feature set performs best
- differences in feature sets not as large for phrase boundaries
Effects of z-scoring

<table>
<thead>
<tr>
<th></th>
<th>non-normalized</th>
<th>normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pitch Accents</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td>83.6</td>
<td>77.0</td>
</tr>
<tr>
<td>Classification</td>
<td>69.0</td>
<td>62.6</td>
</tr>
<tr>
<td><strong>Phrase Boundaries</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td>89.8</td>
<td>83.0</td>
</tr>
<tr>
<td>Classification</td>
<td>87.3</td>
<td>83.2</td>
</tr>
</tbody>
</table>

- speaker-independent experiments using prosody and position features
- the CNN looks or relative changes in speech, and normalizing may lead to a loss in fine differences
Conclusion

- position indicator feature is crucial for this method
- model generalizes well from a speaker-dependent setup to a speaker-independent setting
- presented method can be readily applied to other datasets
- strong and efficient modelling technique that will be used as a basis in future work
- further feature and results analysis necessary
Thank you!

sabrina.stehwien@ims.uni-stuttgart.de
thang.vu@ims.uni-stuttgart.de