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

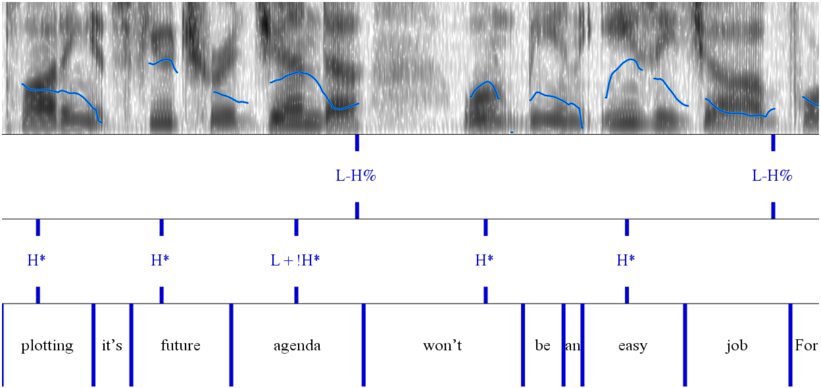
University of Stuttgart
Germany

Institut für
Maschinelle
Sprachverarbeitung

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



L-H%
L-H%
H*
H*
L+!H*
H*
H*
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 ($\rightarrow$ **word-based** recognition)
- address explicit context modelling in a simple and efficient way

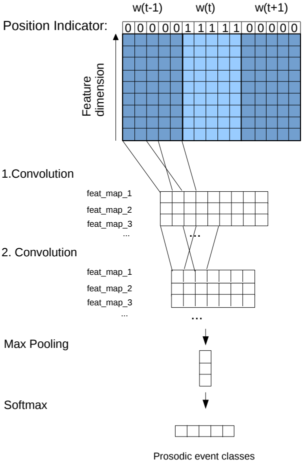
Experimental Focus

- 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

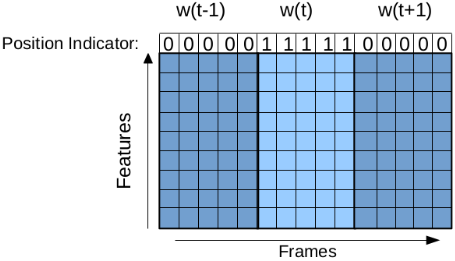
- extracted using the openSMILE toolkit [Eyben et al. 2013]
- two different feature sets:
  - *prosody*: smoothed f0, RMS energy, PCM loudness, voicing probability, Harmonics-to-Noise-Ratio
  - *Mel*: 27 features extraced from the Mel-frequency spectrum
- features computed for each 20ms frame with a 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 4×1
- 2nd layer: 100 kernels of shape 4×1, stride 2×1
- 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

Speakers	f1a	f2b	f3a	m1a	m2b
PA # words	4375	12357	2736	3584	3607
PB # words	4362	12606	2736	5055	3607

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

	one speaker			all speakers		
Feature set	prosody	Mel	pros.+Mel	prosody	Mel	pros.+Mel
Detection						
1 word	84.2	84.2	84.0	81.9	78.3	79.3
3 words	58.3	53.1	53.6	58.2	54.3	55.3
3 words + PF	**86.3**	83.3	83.9	**83.6**	80.3	81.1
Classification						
1 word	74.4	72.7	73.5	68.0	64.7	64.5
3 words	52.4	47.8	47.8	50.5	48.4	48.4
3 words + PF	**76.3**	72.3	72.9	**69.0**	65.9	65.3

all results reported in accuracy (%)

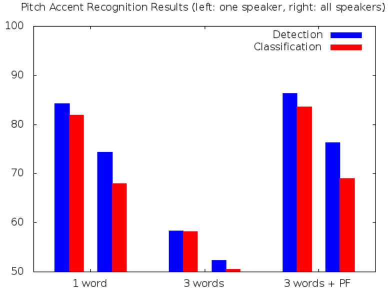
Results: Phrase Boundary Recognition

	one speaker			all speakers		
Feature set	prosody	Mel	pros.+Mel	prosody	Mel	pros.+Mel
Detection						
1 word	87.6	89.2	89.8	86.5	85.3	86.1
3 words	80.3	75.4	75.4	82.7	81.0	80.8
3 words + PF	90.2	90.4	**90.5**	**89.8**	88.3	88.8
Classification						
1 word	85.6	87.6	88.0	85.1	84.4	84.9
3 words	79.7	74.5	74.6	82.5	81.4	81.5
3 words + PF	87.8	88.7	**88.8**	**87.3**	86.2	86.7

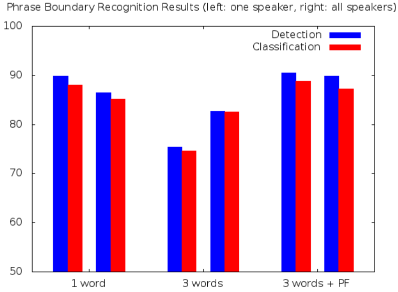
all results reported in accuracy (%)

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

	non-normalized	normalized
Pitch Accents		
Detection	83.6	77.0
Classification	69.0	62.6
Phrase Boundaries		
Detection	89.8	83.0
Classification	87.3	83.2

- 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