Convolutional neural networks can learn duration for detecting pitch accents and lexical stress

Sabrina Stehwien, Antje Schweitzer, Ngoc Thang Vu

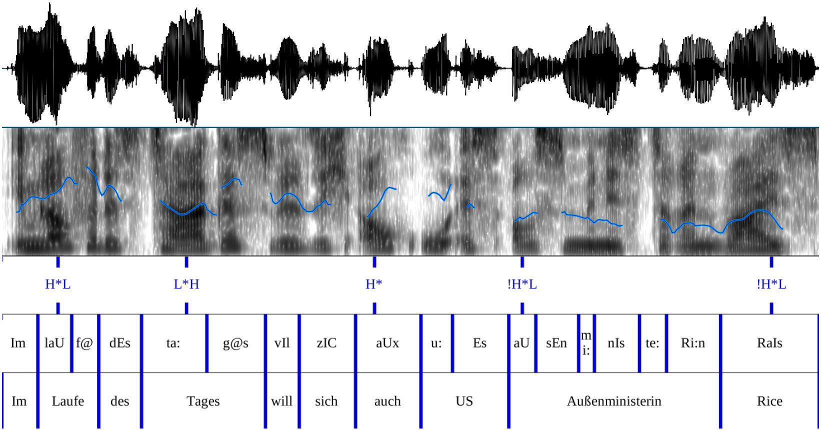
University of Stuttgart
Institute for Natural Language Processing (IMS)

March 6, 2019

Introduction

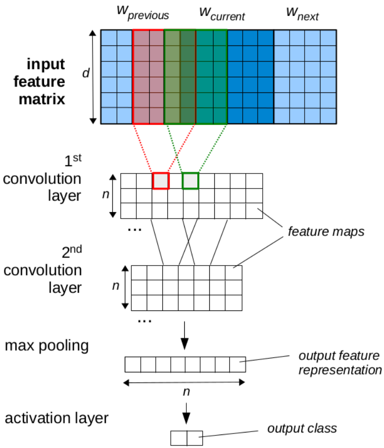
- **pitch accent detection**
  detect which words or syllables are pitch accented
- **lexical stress detection**
  detect which syllables carry (primary) lexical stress
- input: force-aligned speech data (or ASR output)
- similar acoustic features
- **duration** is an important correlate

DIRNDL example



Pitch accent detection with convolutional neural networks

- **advantage:** requires very little preprocessing
- CNN learns high-level feature representation
- input: 3-word input window
- frame-based acoustic features
  - f0, energy, loudness, voicing probability, HtNR, zero-crossing rate
- 2 convolution layers



Stehwien & Vu (2017), Prosodic event recognition using convolutional neural networks with context information. *Proceedings of Interspeech*

Contributions

duration is not an explicit input feature, but provided implicitly by the number of frames for each input word

Experiments test the following **assumptions**:

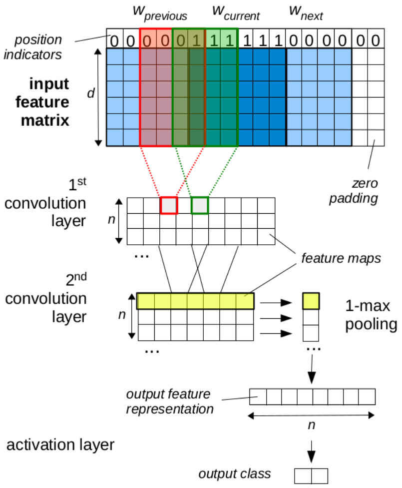
1. the CNN can learn this information on its own
2. method of pooling and padding affects how duration is captured

Tested on several **tasks**:

- pitch accent detection and classification
- word- and syllable-level
- lexical stress detection
- **Data:** DIRNDL German radio news corpus

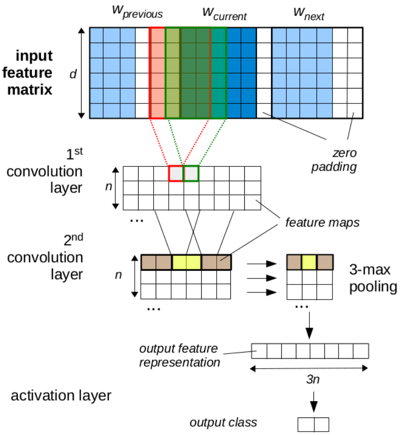
1-max pooling with position indicators

- max pooling selects single neuron with highest activation in each feature map
- zero padding at end of word
- position indicator features mark frames of current word
- kernels span all features
- 100 feature maps
- output feature vector: $n = 100$



3-max pooling

- zero padding at end of each word
- 3 max pooling windows of equal size
- 30 feature maps
- output feature vector: $3n = 90$



Comparison of pooling methods

Compared 1-max to 3-max pooling on word-based pitch accent detection

Results:

- 1-max pooling
  - requires position indicators
  - padding in between words: slight improvement
- 3-max pooling does not require position indicators

setting		+ pos.-ind. (F1)	- pos.-ind. (F1)
1-max pooling	padding at end	**86.8**	63.4
	padding between words	86.2	70.4
3-max pooling	padding between words	85.8	**86.0**

Evidence of duration in CNN output representations

Assumptions: (1) CNN learns duration on its own,
(2) pooling and padding have different effects

Method:

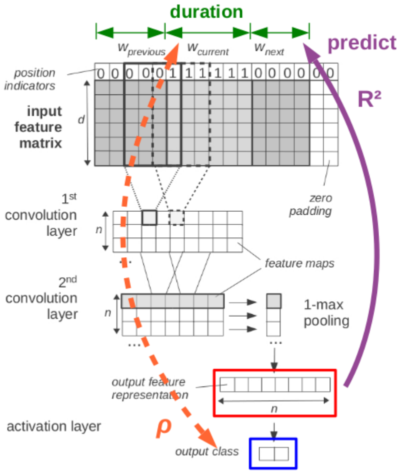
- use CNN output representation to predict duration
- if duration can be approximated $\Rightarrow$ encoded in CNN features
- the better the fit, the more duration information has been learned
- expected to depend on the task
  $\rightarrow$ the more important duration is, the better it should be predicted

Recognition tasks

1. word-based pitch accent detection
2. syllable-based pitch accent detection
3. lexical stress detection (syllables)
4. pitch accent detection on stressed syllables only
5. pitch accent classification (syllables)
   not considering the "none" class!
   - H* vs. H*L vs. L*H
   - H*L vs. L*H

Linear model analysis

- train linear regression model on CNN output
- predict duration of each input word/syllable
- goodness of fit $R^2 \Rightarrow$ estimation of encoded information
- compare correlation between duration and target label (Spearman's ρ)



Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ		R^2			R^2	
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ	R^2			R^2		
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

duration of current word can be approximated
$\Rightarrow$ encoded in CNN output representation

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ	R^2			R^2		
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

correlation between target label and duration explains fit

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ		R^2			R^2	
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

1-max pooling: context word durations not learned well

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ		R^2			R^2	
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

3-max pooling: more context information captured, but current word most important

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ		R^2			R^2	
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

duration more important for pitch accent detection on words than on syllables → correlation with word length

Results: Duration in CNN output representations

task/setting		1-max pooling			3-max pooling		
measure	ρ	R^2			R^2		
duration	w_{cur}	w_{prev}	w_{cur}	w_{next}	w_{prev}	w_{cur}	w_{next}
word-based							
PA detection	0.70	0.11	0.64	0.09	0.48	0.61	0.41
- pos.-ind.		0.06	0.14	0.06			
syllable-based							
PA detection	0.34	0.06	0.42	0.06	0.16	0.40	0.18
stress detection	0.22	0.11	0.31	0.07	0.31	0.38	0.32
PA detection str.-only	0.36	0.09	0.40	0.06	0.24	0.38	0.21
H*/H*L/L*H	-0.04	0.05	0.16	0.12	0.21	0.19	0.15
H*L/L*H	-0.04	0.03	0.14	0.07	0.09	0.17	0.08

no correlation with pitch accent type if presence is given $\rightarrow$ duration more important for detection than type distinction

Summary

- analysis based on linear models provides evidence that certain information is encoded in the CNN output
- **conclusion:** CNN-based pitch accent detector can learn duration on its own from frame-based input
- considered syllable-based tasks and lexical stress detection: correlation with target label matters
- compared 1-max pooling to 3-max pooling
  $\rightarrow$ in both cases, the duration of the current word/syllable is learned best

sabrina.stehwien@ims.uni-stuttgart.de