Universität Stuttgart

www.ims.uni-stuttgart.de

Effects of Adding Word Embeddings to Neural-Network-based Pitch Accent Detection

Sabrina Stehwien, Ngoc Thang Vu, Antje Schweitzer

Abstract

Motivation

- PAD benefits from adding information from text: parts of speech, function vs. content words, word identity
- state-of-the-art deep learning methods use word embeddings to represent syntactic and semantic properties of words
- not previously used for PAD on transcribed speech

Findings

- word embeddings help most when word overlap is significant
- this tends to lead to overfitting → generalization challenging

Model

Required input data

- acoustic signal (.WAV) and transcriptions
- time-aligned at the word level

Convolutional Neural Network

- input matrix: frame-based acoustic features for each trigram
- position features indicate current word
- 2-layer convolutional neural network
  - 1st layer: 100 kernels, size 6 x 7
  - 2nd layer: 100 kernels, size 4 x 2
- dropout: $p = 0.2$, $l2$ regularization

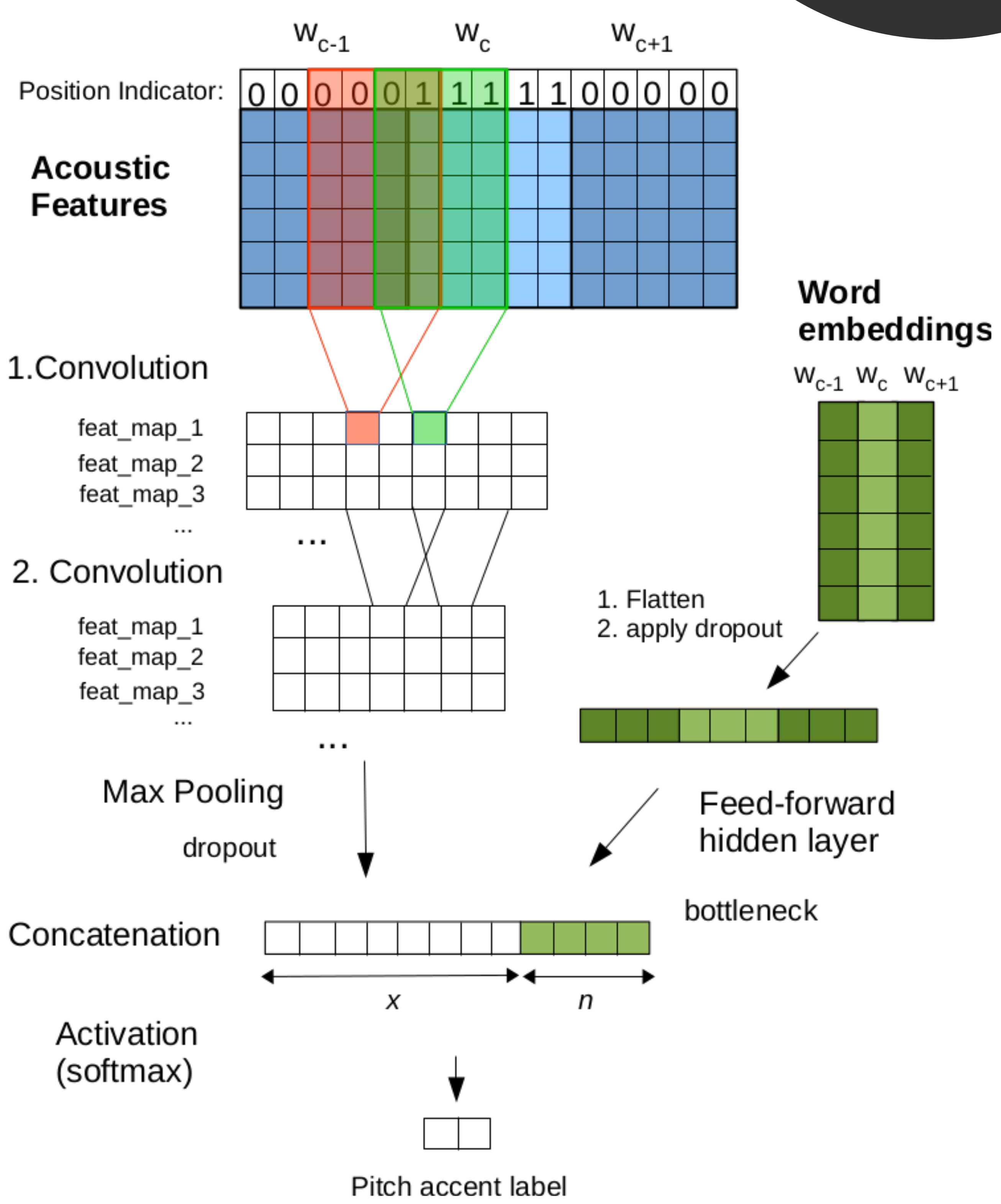
Acoustic Features

6 low-level descriptors extracted using OpenSMILE [1]

RMS energy, loudness*, smoothed F0, voicing probability, harmonics-to-noise-ratio, zero-crossing rate*

Feed-forward Network and Word Embeddings

- input: for each unigram or word in trigram 300-dimensional word embedding vector
- pre-trained word embeddings: *word2vec* [2], *GloVe* [3]
- used as non-trainable matrix weights in hidden layer
- dropout $p = 0.8$, $l2$ regularization
- bottleneck with variable size n



Data

- Boston University Radio News Corpus [4]
  27k words, 51.5% accented
- Boston Directions Corpus (read & spontaneous) [5]
  19k words, 55.5% accented
- LeaP corpus of non-native speech (read & retold stories) [6]
  15k words, 43.1% accented

Experimental Results

Train \ Test	BURNC	BDC	LeaP
BURNC			
acoustic	87.1	74.2	79.2
acoustic+embs	**87.5**	75.5	78.6
embs-only	*78.5*	*71.6*	*76.0*
BDC			
acoustic	82.3	78.0	76.3
acoustic+embs	82.6	**81.2**	77.5
embs-only	*75.0*	*76.0*	*74.5*
LeaP			
acoustic	82.6	72.1	80.5
acoustic+embs	77.7	73.0	**83.5**
embs-only	*67.7*	*68.0*	*80.9*
ALL			
acoustic	86.6	77.4	80.8
acoustic+embs	87.0	80.6	83.4
embs-only	*75.2*	*72.7*	*77.6*

Corpus	BURNC		BDC		LeaP	
Embeddings	glove	w2v	glove	w2v	glove	w2v
unigram						
$n = 10$	87.5	87.6	81.2	80.6	83.5	83.9
$n = 20$	87.4	87.7	81.5	81.1	83.6	83.8
trigram						
$n = 10$	87.7	87.7	82.4	81.1	83.9	83.6
$n = 30$	87.8	87.5	82.7	81.4	83.7	83.8

All results shown in accuracy (%) averaged using 10-fold crossvalidation and 5 repetitions.

Left: within-corpus and cross-corpus experiments using GloVe unigram embeddings, $n = 10$
Below: within-corpus experiments using embeddings with and without context and varying bottleneck sizes

Out-of-vocabulary words and performance on stopwords

- word2vec omits stopwords
  a, and, of, to
- OOVs represented as vector of ones

	BURNC	BDC	LeaP
baseline	98.2	88.9	86.9
GloVe	98.2	92.7	94.2
word2vec	97.8	92.7	94.3

accuracy (%), unigram emb., $n = 10$

	BURNC	BDC	LeaP
GloVe OOV			
tokens	233	19	4
types	64	11	4
accent rate	93%	74%	50%
word2vec OOV			
tokens	3375	2496	1822
types	231	66	6
stopword rate	70.5%	87%	99.9%
accented stopwords	3%	13%	6%
accented remaining	79.5%	83%	100%

Acknowledgements

This work was funded by the German Research Foundation DFG (SFB 732, A8).

[1] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in opensmile, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at ICLR*, 2013.

[3] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[4] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University, 1995.

[5] J. Hirschberg and C. H. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *34th annual meeting of the ACL*, 1996.

[6] J.-T. Milde and U. Gut. A prosodic corpus of non-native speech. In *Speech Prosody*, 2002.