

# Automatic classification of lexical stress errors for German CAPT

Anjana Vakil and Jürgen Trouvain



UNIVERSITÄT  
DES  
SAARLANDES

Department of Computational Linguistics and Phonetics  
Saarland University, Saarbrücken, Germany

SLaTE 2015, Leipzig  
4 September 2015

Accentuation/prominence of syllable(s) in a word

In German:

- ▶ Variable placement, contrastive function

um·FAHR·en	vs.	UM·fahr·en
<i>to drive around</i>		<i>to run over</i>

- ▶ Reflected by duration, fundamental frequency (F0), intensity<sup>1</sup>
- ▶ Impacts intelligibility of non-native (L2) speech<sup>2</sup>

---

<sup>1</sup>Dogil and Williams 1999.

<sup>2</sup>Hirschfeld 1994.

- ▶ Contrastive LS notoriously difficult for French speakers<sup>1</sup>
- ▶ CAPT offers huge potential for individualized instruction
- ▶ Classification of LS errors in L2 German unexplored
- ▶ Promising recent work using machine learning for classification of English stress patterns<sup>2</sup>

**Our goal:** explore classification-based detection of lexical stress errors by French learners of German

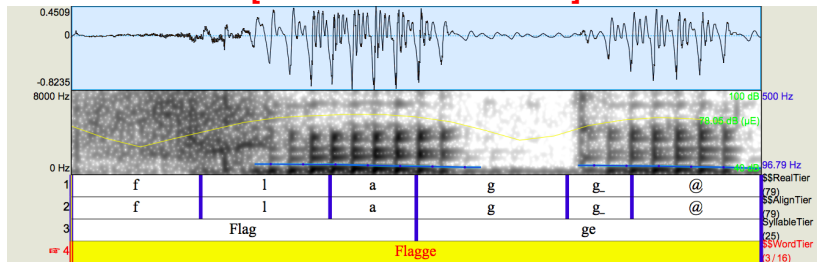
---

<sup>1</sup>Dupoux et al. 1997.

<sup>2</sup>Kim and Beutnagel 2011; Shahin et al. 2012.

# Subset of IFCASL corpus of French-German speech<sup>1</sup>

[TODO new screenshot]



Extracted utterances of 12 bisyllabic, initial-stress words

- ▶ 668 tokens from 56 French speakers - manually annotated
- ▶ 477 tokens from 40 German speakers - assumed correct

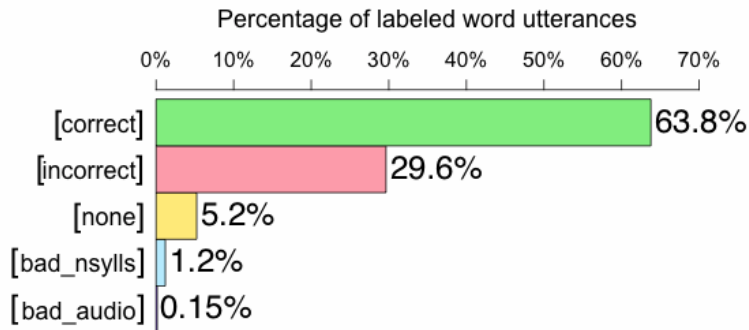
<sup>1</sup>Fauth et al. 2014.

- ▶ Each token assigned a class label:
  - 3 stress classes: [correct], [incorrect], [none]
  - 2 error classes: [bad\_nsylls], [bad\_audio]
- ▶ 15 annotators (12 native), each token labeled by  $\geq 2$

Overall pairwise inter-annotator agreement

	Mean	Maximum	Median	Minimum
% Agreement	54.92%	83.93%	55.36%	23.21%
Cohen's $\kappa$	0.23	0.61	0.26	-0.01

- ▶ Variability not explained by L1 or expertise
- ▶ Single gold-standard label selected for each token



Train & evaluate CART classifiers using WEKA toolkit<sup>1</sup>

## Training data

- ▶ Manually annotated L2 utterances
- ▶ Automatically annotated L1 utterances (all [correct])

## Held-out testing data

- ▶ Feature comparison: 1/10 of L2 utterances (random)
- ▶ Unseen speakers: all utterances from 1 of 56 L2 speakers

## Evaluation

- ▶ Compute agreement (% and  $\kappa$ ) with gold standard
- ▶ Average across 10 or 56 folds

---

<sup>1</sup>[www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

## Prosodic feature sets

- ▶ DUR - Duration (relative syllable & nucleus lengths)
- ▶ F0 - Fundamental frequency (mean, max., min., range)
- ▶ INT - Intensity (mean, max.)

Pitch and energy contours calculated using JSnoori software<sup>1</sup>

For German stress, duration seemingly best indicator, then F0<sup>2</sup>

---

<sup>1</sup>`jsnoori.loria.fr`

<sup>2</sup>Dogil and Williams 1999.

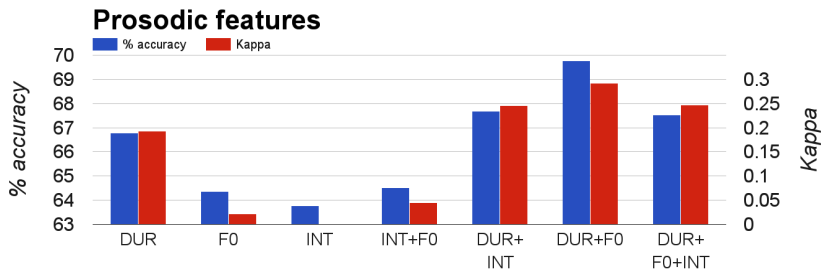


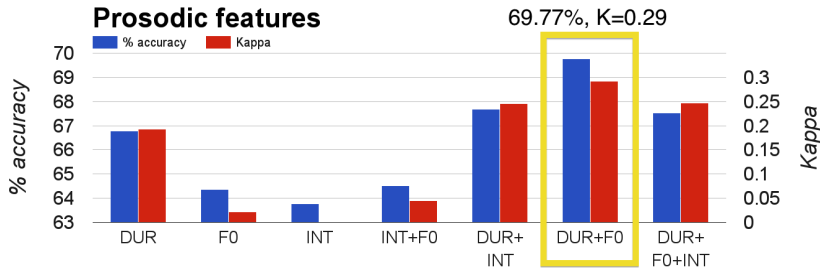
## Prosodic feature sets

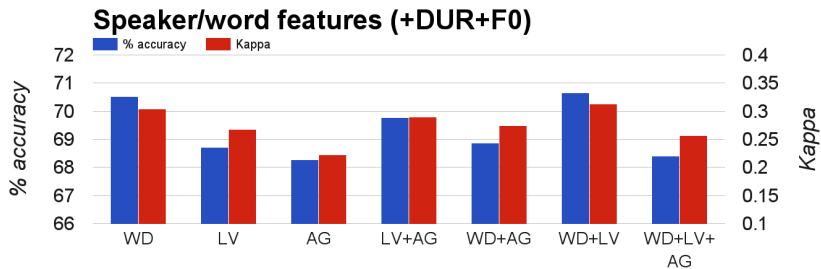
- ▶ DUR - Duration
- ▶ F0 - Fundamental frequency
- ▶ INT - Intensity

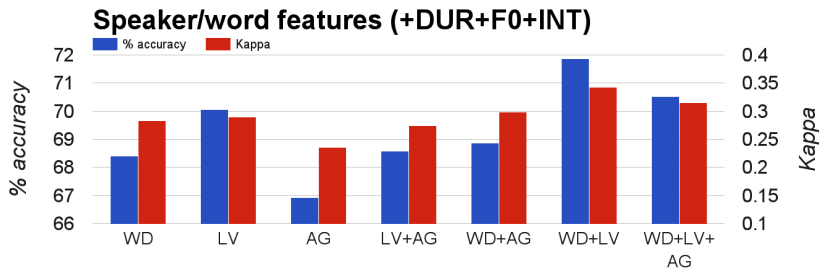
## Other features

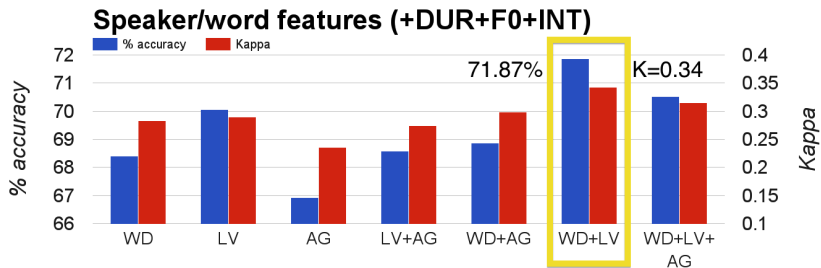
- ▶ WD - Word uttered (e.g. *Flagge*)
- ▶ LV - Speaker's skill level (A2|B1|B2|C1)
- ▶ AG - Speaker's age/gender (Girl|Boy|Woman|Man)

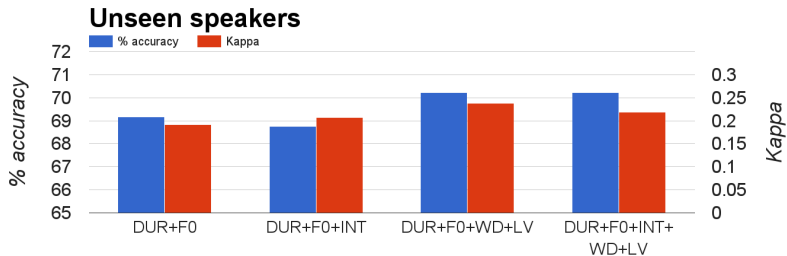


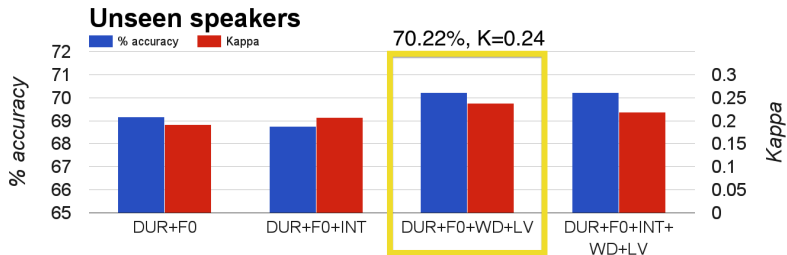














	% agreement	$\kappa$
Best classifier vs. gold standard	71.87%	0.34
Majority baseline vs. gold standard	63.77%	0.00
Human vs. human	54.92%	0.23

- ▶ Results are encouraging in this context
- ▶ Still want better performance for real-world use

- ▶ Classification-based diagnosis of lexical stress errors  
novel approach in German CAPT
- ▶ Results of  $>70\%$  accuracy encouraging  
(especially considering low human-human agreement)
- ▶ Still much room for improvement

## Future directions

- ▶ More powerful machine learning algorithms
- ▶ Additional features (e.g. vowel quality, phrase information)
- ▶ Online, semi-supervised learning/active learning

- ▶ G. Dogil and B. Williams. “The phonetic manifestation of word stress”. In: *Word Prosodic Systems in the Languages of Europe*. Ed. by H. van der Hulst. Walter de Gruyter, 1999. Chap. 5, pp. 273–334.
- ▶ E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler. “A Destressing ‘Deafness’ in French?” In: *Journal of Memory and Language* 36.3 (Apr. 1997), pp. 406–421.
- ▶ C. Fauth, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouviet, J. Jügler, Y. Laprie, O. Mella, and B. Möbius. “Designing a Bilingual Speech Corpus for French and German Language Learners: A Two-Step Process”. In: *9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland, 2014, pp. 1477–1482.
- ▶ U. Hirschfeld. *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Forum Phonetikum. 1994.
- ▶ Y.-J. Kim and M. C. Beutnagel. “Automatic assessment of American English lexical stress using machine learning algorithms”. In: *SLaTE*. 2011, pp. 93–96.
- ▶ M. A. Shahin, B. Ahmed, and K. J. Ballard. “Automatic classification of unequal lexical stress patterns using machine learning algorithms”. In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, Dec. 2012, pp. 388–391.