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

Lexical stress [TODO (LS)] in German

Accentuation/prominence of syllable(s) in a word

In German:

- Variable placement, contrastive function

  um·FAHR·en vs. UM·fahr·en
  to drive around *to run over*

- Reflected by duration, fundamental frequency (F0), intensity[1]
- Impacts intelligibility of non-native (L2) speech[2]

[1] Dogil and Williams 1999.
[2] Hirschfeld 1994.

CAPT for lexical stress errors [TODO (LSEs)]

UNIVERSITÄT DES SAARLANDES

- Contrastive LS notoriously difficult for French speakers[1]
- 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[2]

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

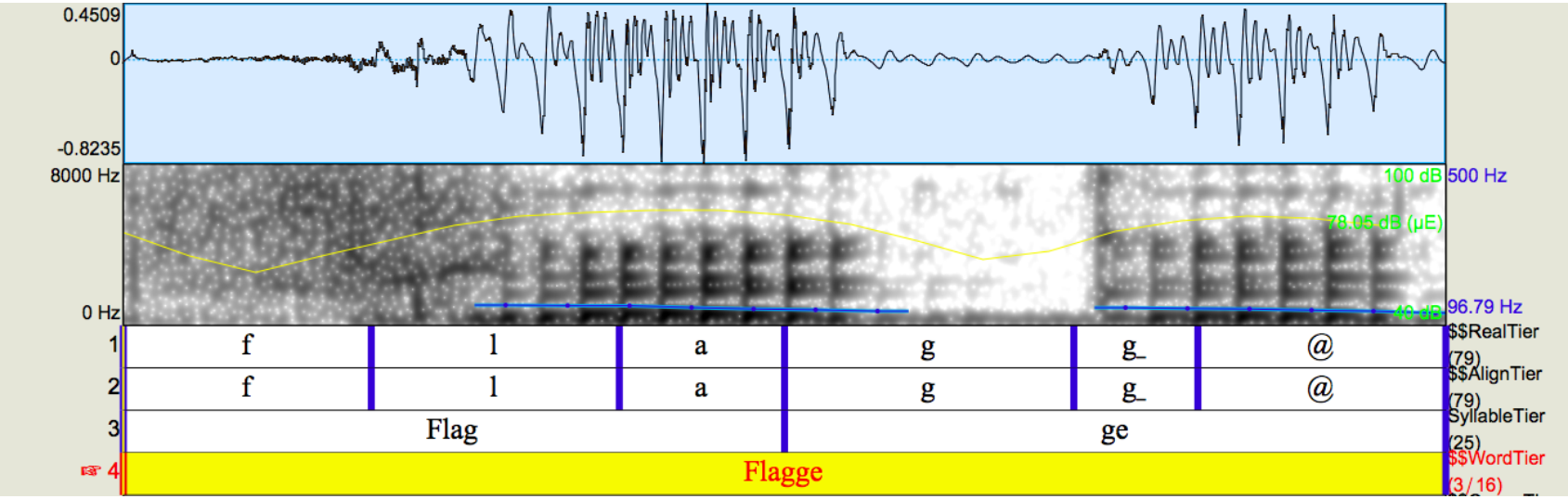
[1] Dupoux et al. 1997.

[2] Kim and Beutnagel 2011; Shahin et al. 2012.

Data

Subset of IFCASL corpus of French-German speech[1]

[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

[1] Fauth et al. 2014.

Data annotation

- Each token assigned a class label:

  3 stress classes:
  [correct], [incorrect], [none]

  2 error classes:
  [bad_nsylls], [bad_audio]

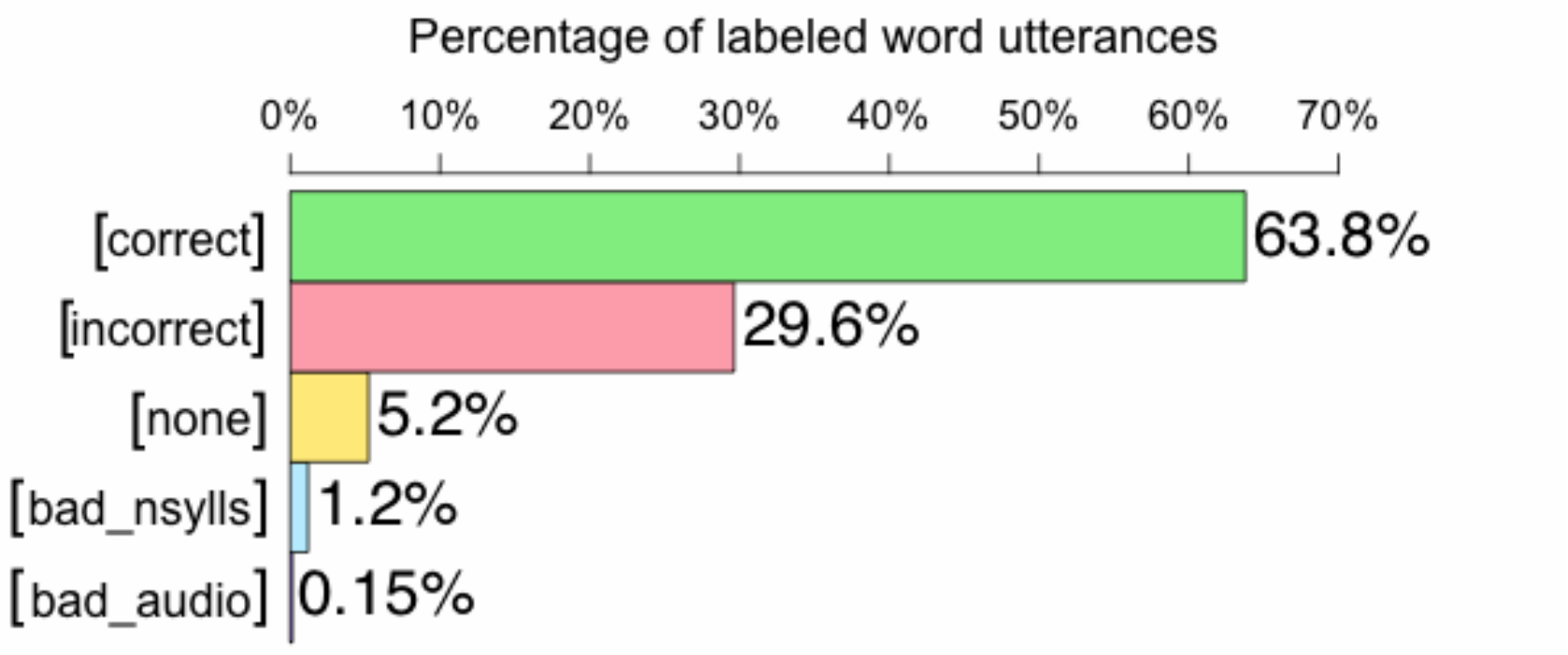
- 15 annotators (12 native), each token labeled by $\geq$2

Overall pairwise inter-annotator agreement

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

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

Data annotation results



Method

Train & evaluate CART classifiers using WEKA toolkit[1]

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 κ) with gold standard
- Average across 10 or 56 folds

[1] `www.cs.waikato.ac.nz/ml/weka`

Feature sets

UNIVERSITÄT DES SAARLANDES

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[1]

For German stress, duration seemingly best indicator, then F0[2]

[1] `jsnoori.loria.fr`

[2] Dogil and Williams 1999.

Feature sets

UNIVERSITÄT DES SAARLANDES

Prosodic feature sets

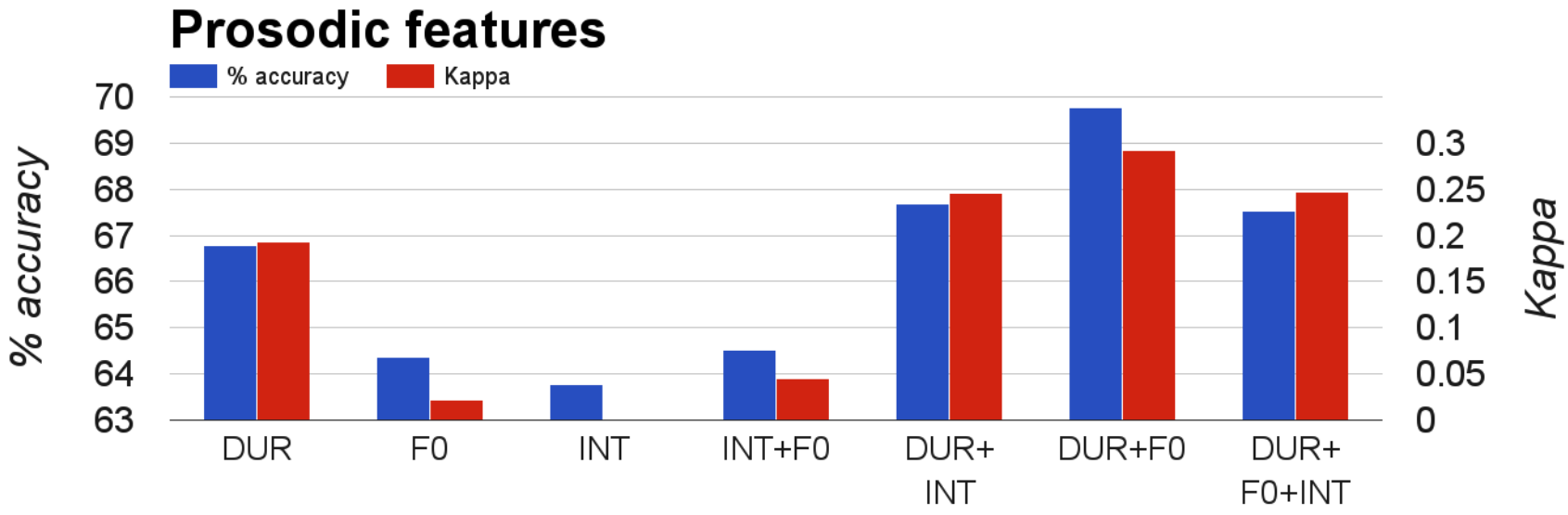
- DUR - Duration
- F0 - Fundamental frequency
- INT - Intensity

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)

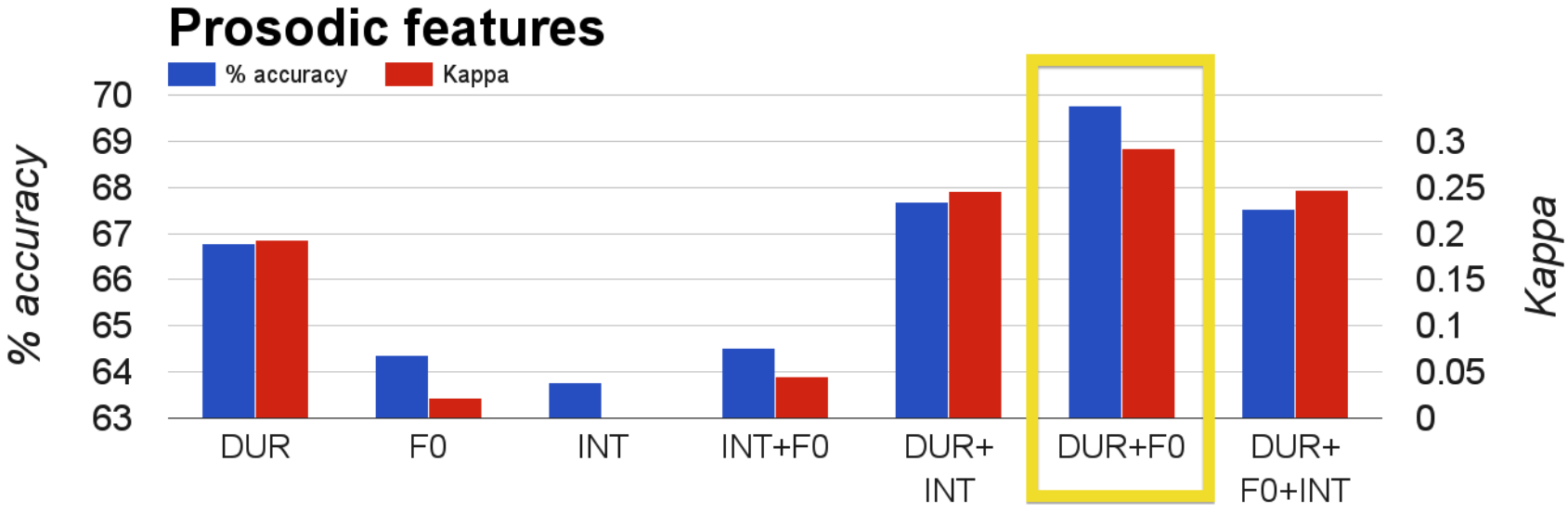
Results

How well can lexical stress errors be classified?



Results

How well can lexical stress errors be classified?

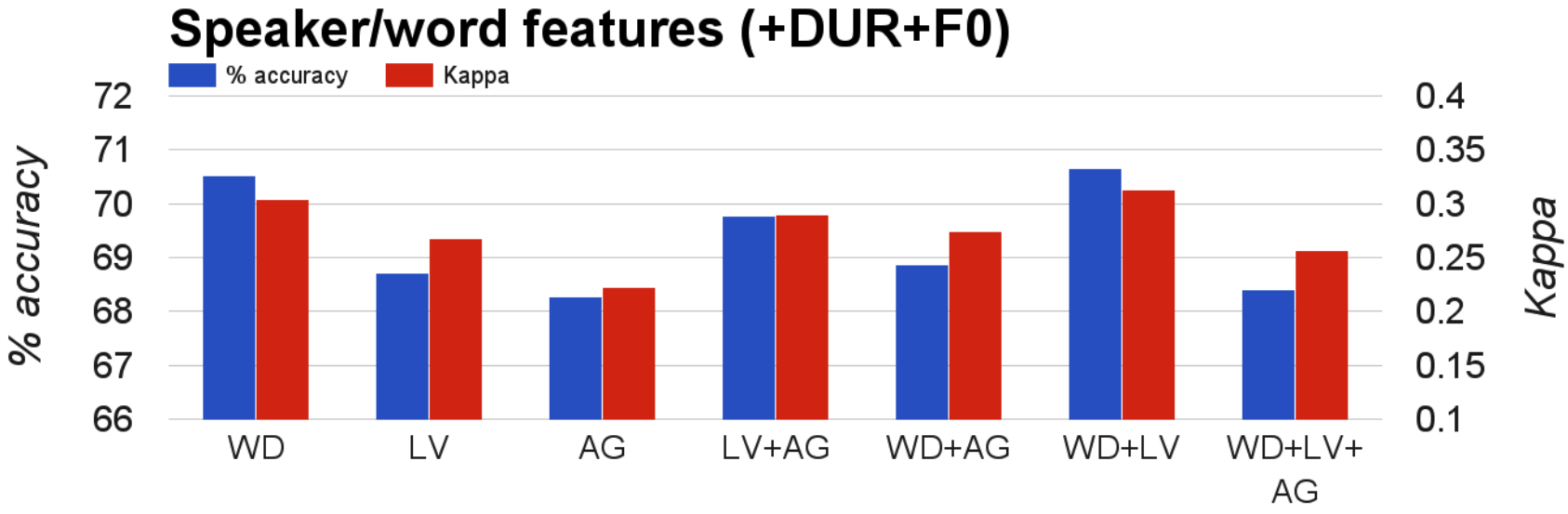


Best performance using only prosodic features: DUR+F0

- % Accuracy: 69.77%
- κ: 0.29

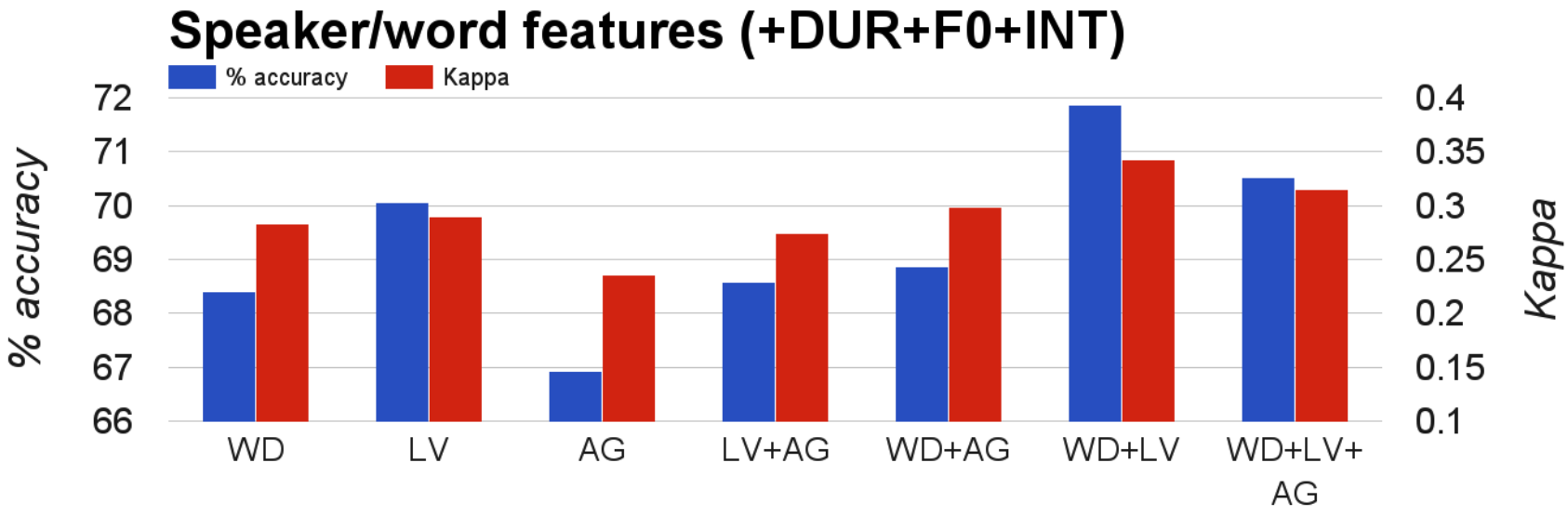
Results

How well can lexical stress errors be classified?



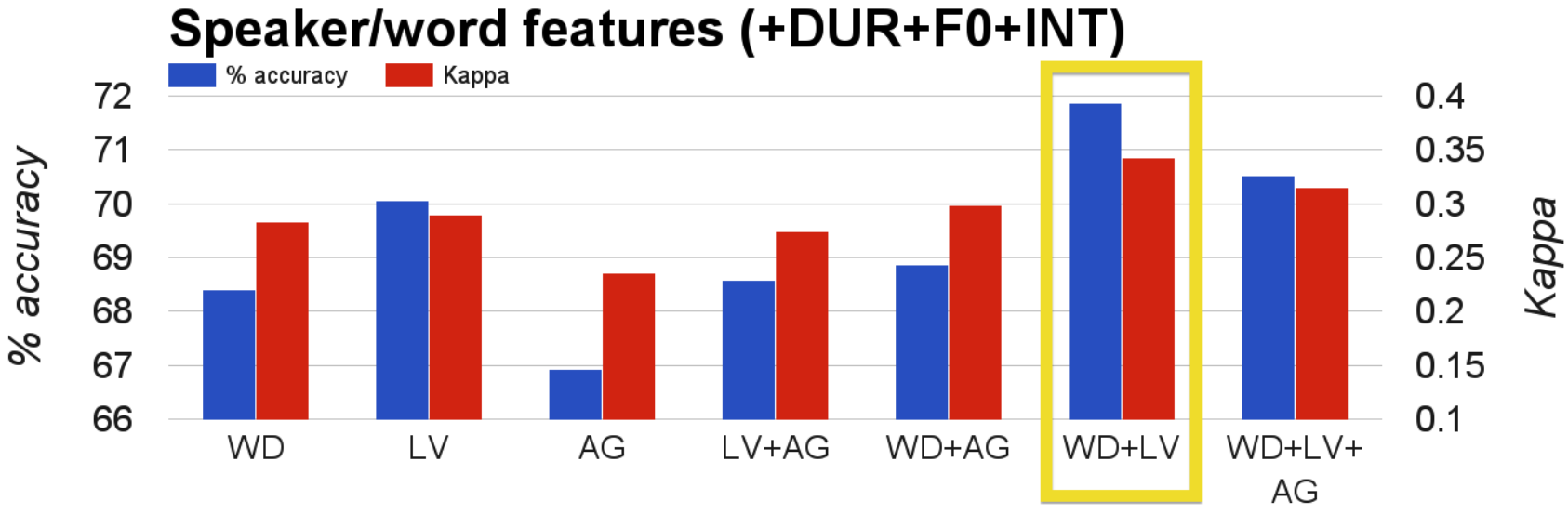
Results

How well can lexical stress errors be classified?



Results

How well can lexical stress errors be classified?



Best performance overall: WD+LV+DUR+F0+INT

- % Accuracy: 71.87%
- κ: 0.34

Results

UNIVERSITÄT DES SAARLANDES

Unseen speakers

Results

How does classification accuracy compare with human agreement?

	% agreement	κ
Best classifier vs. gold standard	71.87%	0.34
Mean human vs. human	54.92%	0.23

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

Conclusions

Selected references

- 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. Jouvet, 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 Phoneticum. 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.