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: 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, F0, intensity
- Impacts intelligibility of non-native (L2) speech

Background & motivation

- Contrastive lexical stress (LS) difficult for French speakers
- CAPT can help; requires automatic diagnosis
- Classification of LS errors in L2 German unexplored

Classification of LS errors by French learners of German

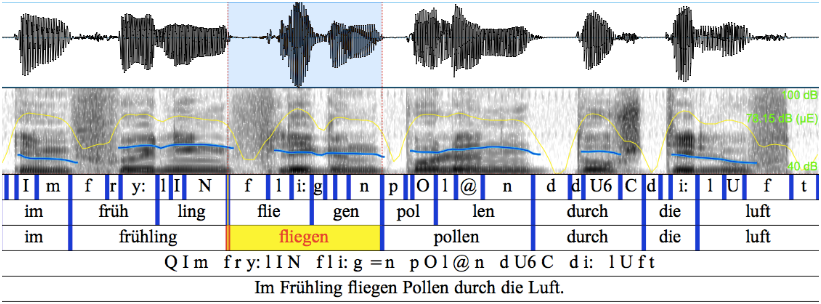
How feasible is it?

Which features are most useful?

Data

UNIVERSITÄT DES SAARLANDES

Subset of IFCASL corpus of French-German speech (Fauth et al. 2014)



Extracted utterances of 12 bisyllabic, initial-stress words

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

Data annotation

- Each token assigned a class label:

[correct], [incorrect], [none]

[bad_nsylls], [bad_audio]

- 15 annotators (12 native), each token labeled by $\geq$2
- Varying phonetics/phonology expertise

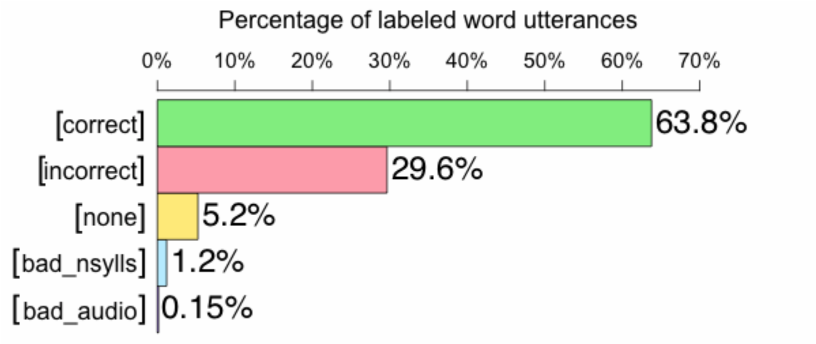
Data annotation

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 annotator L1 or expertise
- Single gold-standard label selected for each token

Data annotation results



Method

Train & evaluate CART classifiers using WEKA toolkit

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
- Cross-validation (10 or 56 folds)

Feature sets

UNIVERSITÄT DES SAARLANDES

Prosodic feature sets

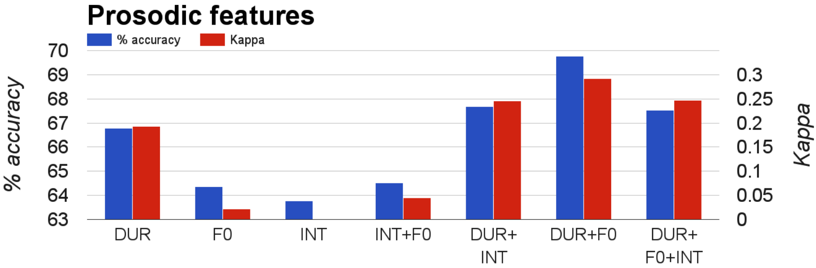
- DUR - Duration (relative syllable & nucleus lengths)
- F0 - Fundamental frequency (mean, max., min., range)
- INT - Intensity (mean, max.)

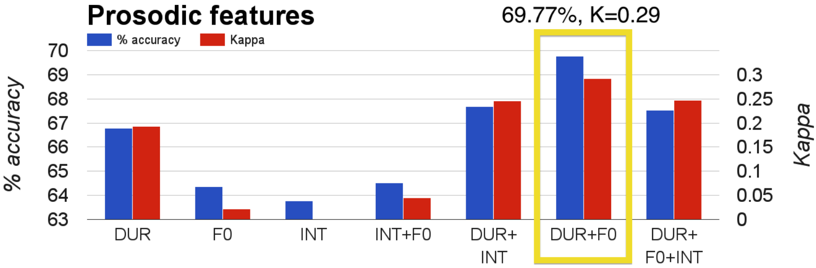
Pitch and energy contours calculated using JSnoori software (`http://jsnoori.loria.fr`)

Other features

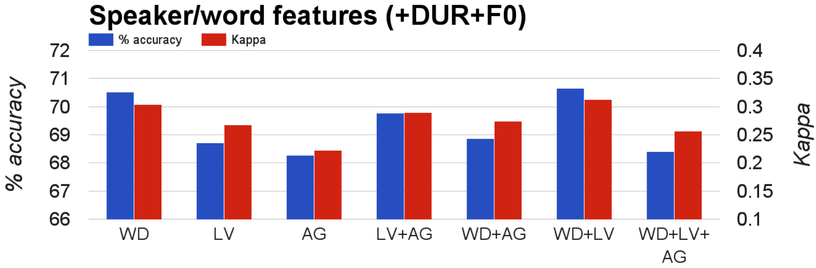
- WD - Word uttered (e.g. *Flagge*)
- LV - Speaker's CEFR skill level (A2|B1|B2|C1)
- AG - Speaker's age/gender (Girl|Boy|Woman|Man)

Results


Prosodic features
% accuracy
Kappa
70
69
68
67
66
65
64
63
% accuracy
0.3
0.25
0.2
0.15
0.1
0.05
0
Kappa
DUR
F0
INT
INT+F0
DUR+ INT
DUR+F0
DUR+ F0+INT

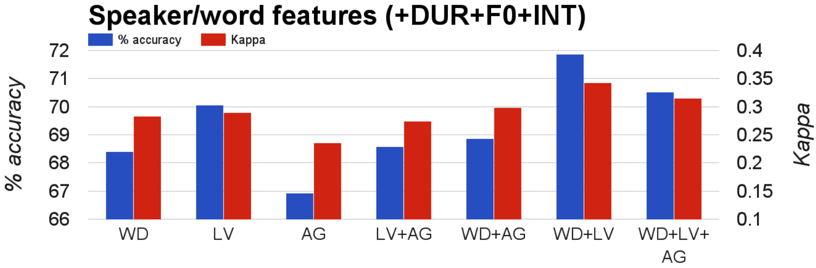

Prosodic features
69.77%, K=0.29
% accuracy
Kappa
70
69
68
67
66
65
64
63
0.3
0.25
0.2
0.15
0.1
0.05
0
DUR
F0
INT
INT+F0
DUR+ INT
DUR+F0
DUR+ F0+INT

Results

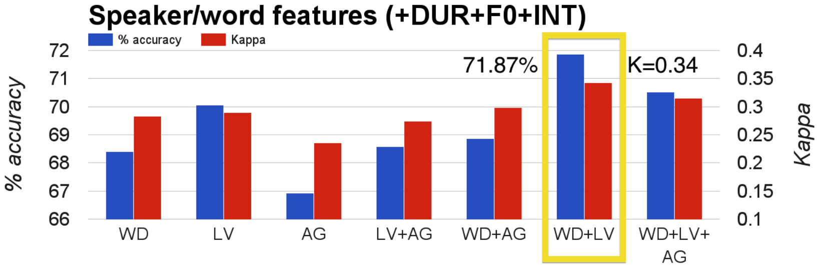


Speaker/word features (+DUR+F0)
% accuracy
Kappa
72
71
70
69
68
67
66
0.4
0.35
0.3
0.25
0.2
0.15
0.1
WD
LV
AG
LV+AG
WD+AG
WD+LV
WD+LV+AG

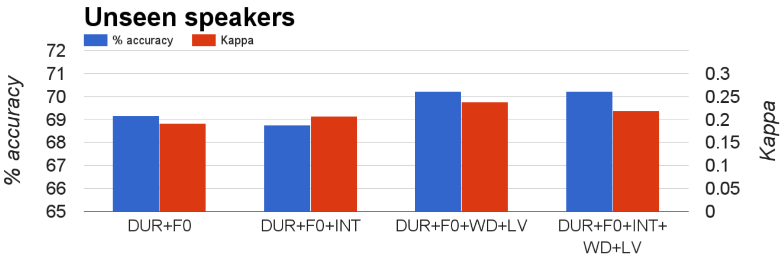
Results



Speaker/word features (+DUR+F0+INT)
% accuracy
Kappa
72
71
70
69
68
67
66
% accuracy
0.4
0.35
0.3
0.25
0.2
0.15
0.1
Kappa
WD
LV
AG
LV+AG
WD+AG
WD+LV
WD+LV+AG

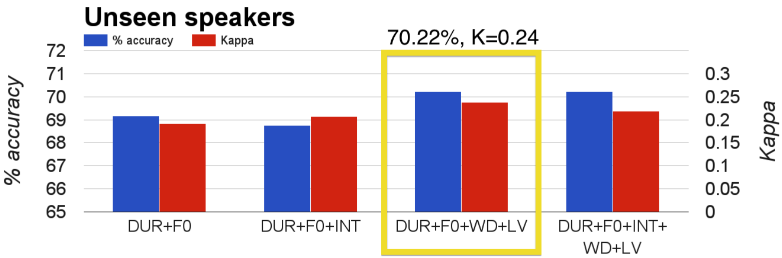

Speaker/word features (+DUR+F0+INT)
% accuracy
Kappa
72
71
70
69
68
67
66
0.4
0.35
0.3
0.25
0.2
0.15
0.1
71.87%
K=0.34
WD
LV
AG
LV+AG
WD+AG
WD+LV
WD+LV+AG
% accuracy
Kappa

Results



Unseen speakers
% accuracy
Kappa
% accuracy
72
71
70
69
68
67
66
65
0.3
0.25
0.2
0.15
0.1
0.05
0
Kappa
DUR+F0
DUR+F0+INT
DUR+F0+WD+LV
DUR+F0+INT+ WD+LV

Results


Unseen speakers
% accuracy
Kappa
70.22%, K=0.24
72
71
70
69
68
67
66
65
% accuracy
0.3
0.25
0.2
0.15
0.1
0.05
0
Kappa
DUR+F0
DUR+F0+INT
DUR+F0+WD+LV
DUR+F0+INT+
WD+LV

	% agreement	κ
Best classifier vs. gold standard		
Random test set	71.87%	0.34
Unseen speakers	70.22%	0.24
Majority ([correct]) classifier vs. gold	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

Conclusion

UNIVERSITÄT DES SAARLANDES

- 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