Automatic classification of lexical stress errors for German CAPT

Anjana Sofia Vakil, Jürgen Trouvain

Department of Computational Linguistics & Phonetics
Saarland University, Saarbrücken, Germany
[anjanav,trouvain]@coli.uni-saarland.de

Abstract

Lexical stress plays an important role in the prosody of German, and presents a considerable challenge to native speakers of languages such as French who are learning German as a foreign language. These learners stand to benefit greatly from Computer-Assisted Pronunciation Training (CAPT) systems which can offer individualized corrective feedback on such errors, and reliable automatic detection of these errors is a prerequisite for developing such systems. With this motivation, this paper presents an exploration of the use of machine learning methods to classify non-native German lexical stress errors. In classification experiments using a manually-annotated corpus of German word utterances by native French speakers, the highest observed agreement between the classifier's output and the gold-standard labels exceeded the inter-annotator agreement between humans asked to classify lexical stress errors in the same data. These results establish the viability of classification-based diagnosis of lexical stress errors for German CAPT.
Index Terms: CAPT, prosody, German

1. Introduction

For adult learners of a second language (L2), the phonological system of the L2 can pose a variety of difficulties. For certain L2s, such as German or English, one important difficulty involves the accurate prosodic realization of lexical stress, i.e. the accentuation of certain syllable(s) in a given word, with the placement of stress within a word varying freely and carrying a contrastive function in such languages [1]. Lexical stress is an important part of German word prosody, one which impacts the intelligibility of non-native German speech [2]. Coping with this phenomenon in German is especially challenging for native (L1) French speakers, because lexical stress is realized very differently (or perhaps not at all) in the French language [3].

To overcome this difficulty and improve their L2 word prosody, learners generally need individualized attention from a language instructor; however, the lack of attention typically given to pronunciation in the foreign language classroom, along with other factors such as high student-to-teacher ratios, often make this unfeasible [4, 5]. Fortunately, advances in Computer-Assisted Pronunciation Training (CAPT) over recent decades have made it possible to automatically provide highly individualized analysis of learners' prosodic errors, as well as corrective feedback, and thus to help learners achieve more intelligible pronunciation in the target language.

This paper describes work that advances the state of German CAPT by applying machine learning methods to the task of diagnosing lexical stress errors in non-native German speech, a necessary prerequisite for delivering individualized corrective feedback on such errors in a CAPT system. The paper is organized as follows: Section 2 provides background on the phenomenon of lexical stress as it is realized in German and French word prosody, motivates this work's focus on lexical stress errors, and summarizes related past work. Section 3 describes the manual annotation of lexical stress errors in a small corpus of L2 German speech, i.e. the creation of labeled training and test data for the classification experiments described in section 4. Section 5 presents and analyzes the results of these experiments. Finally, section 6 offers some concluding remarks and possible directions for future work.

2. Background and related work

Broadly speaking, lexical stress is the phenomenon of how a given syllable is accentuated within a word [1], such that this syllable is perceived as "standing out" [6]. This perceived prominence of a syllable is reflected by the prosodic parameters duration, fundamental frequency (F0) and intensity.

In variable-stress languages, such as German and English, the location of lexical stress in a word is not always predictable, so knowing a word requires, in part, knowing its stress pattern. This allows lexical stress to serve a contrastive function in these languages, e.g. distinguishing *UMfahren* (to run over with a car) from *umFAHRen* (to drive around) in German. However, in other languages stress is fixed, i.e. always falls on a certain position in the word (e.g. the final syllable).While French has often been categorized as a fixed-stress language, given that word-final syllables are made prominent (lengthened) when a word is pronounced in isolation, some argue that it may be more properly considered a language without lexical stress, in that speakers do not seem to accentuate any syllable within the word, with word-final lengthening effects explained by interactions with the realization of phrasal accent (lengthening of the final syllable in each prosodic group or phrase) [3, 7]. Regardless, French has no contrastive word-level stress and in this respect differs considerably from German.

Although little research has been done on the nature of lexical stress errors in German spoken by French natives, Hirschfeld and Trouvain [5] report that such errors are commonly observed in this particular L1-L2 pair. Studies on French speakers of Spanish, another contrastive-stress language, have revealed these speakers to be seemingly "deaf" to lexical stress, i.e. to have significant and lasting difficulties perceiving and remembering stress contrasts [7]. With respect to production, studies of French learners of Dutch [3] and English [8] have also shown that these speakers frequently make lexical stress errors, and tend to (incorrectly) stress word-final syllables.

Lexical stress errors may also have a high impact on L2 intelligibility, which is generally considered the most important goal of pronunciation training, as opposed to lack of a "foreign accent" [9, 10]. Prosodic errors have often been found to have a larger impact on the perceived intelligibility of L2 speakers than

segmental errors, and lexical stress errors may have a particularly strong impact on intelligibility in variable-stress languages like English and Dutch [1, 10]. Relatively little research has been done on how various pronunciation errors affect intelligibility in L2 German specifically, but some studies suggest that lexical stress errors may hinder intelligibility more than other error types [2, 5].

The frequency and impact of lexical stress errors by French speakers of German thus motivate the development of German CAPT tools focusing on such errors. However, the feasibility of reliable automatic detection of this type of L2 German error remains to be investigated. To our knowledge, no work has been reported on automatic classification-based diagnosis of such errors in L2 German speech, but in recent years machine learning methods have been applied with apparent success to the classification of lexical stress patterns in English. Kim and Beutnagel [11] experimented with various algorithms to classify stress patterns in high-quality recordings of 3- and 4-syllable English words uttered by L1 speakers, reporting accuracy in the 80-90% range; in pilot experiments with low-quality recordings, however, they report lower accuracy: 70-80% on L1 speech and only 50-60% on L2 speech. Shahin et al. [12] trained Neural Networks to classify stress patterns in bisyllabic words uttered by L1 English children, and reported classification accuracy over 90% for some patterns. Building on these related investigations, this paper explores automatic classification-based diagnosis of German lexical stress errors, with a particular focus on those made by L1 French speakers.

3. Data

Error-annotated speech data from German learners is a prerequisite for the supervised training and evaluation of classifiers for lexical stress realizations in L2 German speech, yet to our knowledge no corpus of learner German with such annotation is publicly available. To fill this need, as well as to shed light on the perception of lexical stress errors in L2 German speech, a small corpus of speech by L1 French learners of German was manually annotated for such errors.

3.1. The IFCASL corpus of learner speech

The learner speech data used in this work has been excerpted from the IFCASL corpus [13], a collection of phonetically diverse utterances in French and German spoken by both native speakers and non-native speakers with the other language as L1. The corpus contains recordings of approximately 50 L1 speakers of each language reading sentences (and a short text) in both languages, such that both L1 and L2 speech was recorded for each speaker. Each L1 speaker group has an even gender distribution, and contains approximately 10 children (adolescents of 15-16 years of age) and 40 adults. A variety of L2 proficiency levels are also represented in the corpus: adults span CEFR[1] levels A2 (beginner) through C1 (advanced), children levels A2 (beginner) and B1 (low intermediate).

The annotation effort described here focuses exclusively on the German-language subset of the corpus. Only utterances from the sub-corpus of L2 German speech by L1 French speakers (henceforth IFCASL-FG) were manually annotated; native utterances from the L1 German sub-corpus (IFCASL-GG) were assumed to contain only correct lexical stress realizations.

[1]Common European Framework of Reference for Languages, www.coe.int/lang-CEFR

Table 1: Word types annotated for lexical stress errors. Canonical pronunciations are given in IPA notation. The rightmost column lists the number of tokens (utterances) of each word type in the dataset.

Word type	Pronunciation	Part of speech	English meaning	Tokens
E-mail	/ˈiː.meɪl/	noun	e-mail	56
Flagge	/ˈfla.gə/	noun	flag	55
fliegen	/ˈfliː.gn̩/	verb	to fly	56
Frühling	/ˈfryː.lŋ/	noun	spring (season)	56
halten	/ˈhal.tn̩/	verb	to hold	56
manche	/ˈman.çə/	pronoun	some	56
Mörder	/ˈmœɐ.dɐ/	noun	murderer	56
Pollen	/ˈpɔ.lən/	noun	pollen	55
Ringen	/ˈʁɪŋ.ən/	noun	rings	55
Tatort	/ˈtaːt.ʔɔɐt/	noun	crime scene	56
tragen	/ˈtʁaː.gn̩/	verb	to wear	55
Tschechen	/ˈtʃɛ.çn̩/	noun	Czechs	56

In addition to the recordings themselves, the IFCASL corpus contains phone- and word-level segmentations of each utterance, produced automatically by forced alignment [13]. Although the corpus also contains manual corrections of these segmentations, the work reported here relies exclusively on the automatic segmentations to mimic the conditions of a fully automatic CAPT system. As the corpus does not include syllable-level segmentations, we created these for each annotated utterance automatically, based on the phone segmentations.

The subset of IFCASL-FG selected for manual error annotation ("the dataset") consists of utterances of twelve bisyllabic word types (see table 1), each of which has primary stress on the initial syllable. Only bisyllabic words were selected to simplify comparison between stressed and unstressed syllables, and only initial-stress words because this is the stress pattern which native (L1) French speakers are expected to have the most difficulty producing in German (see section 2).

Using the automatic segmentations, tokens (utterances) of each selected word type were extracted from the recorded sentences. The dataset comprises 668 word tokens in total; token counts for each word type are listed in table 1.

3.2. Annotation method

The annotation task consisted of assigning one of the following labels to each word token (utterance) in the dataset:

- [correct]: the correct (initial) syllable was clearly stressed
- [incorrect]: the incorrect (final) syllable was clearly stressed
- [none]: neither syllable was clearly stressed, or the annotator was unable to determine which syllable was stressed
- [bad_nsylls]: syllable insertion(s) or deletion(s) interfered with the annotator's ability to judge the stress realization
- [bad_audio]: technical problem(s) (e.g. noise, inaccurate segmentation) interfered with the annotator's judgment

Annotation was performed using a graphical tool, which displayed the given word's text, and allowed the annotator to

Table 2: Overall pairwise agreement between annotators

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

listen to the given word utterance, as well as the sentence utterance from which it was extracted, as many times as they wished. The annotator then clicked one of five buttons, corresponding to the possible labels, to record their judgment. A single annotation session consisted of annotating all 55-56 tokens of each of three word types, and lasted approximately 15 minutes.

A total of 15 annotators participated, varying with respect to their L1 and level of phonetics/phonology expertise. The native languages represented included German (12 annotators), English (2), and Hebrew (1); the L1 English and Hebrew speakers all speak L2 German. In terms of expertise, the annotators were broadly categorized as *experts* (professional phonetics/phonology researchers), *intermediates* (university students enrolled in an experimental phonology course), or *novices* (those with negligible phonetics/phonology training or experience annotating speech data). Among the 15 annotators, there were two experts, 10 intermediates, and three novices. Non-experts were included in the annotator group because the ultimate goal is successful L2 German communication, and it can be assumed that in the vast majority of cases learners will be communicating with non-experts; therefore, it is important that the perception of errors by non-experts not be ignored in favor of experts' perception.

Each annotator was assigned three word types to annotate in a single session, with the exception of one who annotated six word types over two sessions. Assignments ensured that each word token was annotated by at least two native German speakers, and maximized the amount of overlap between annotators in order to obtain as many pairwise measures of annotator agreement as possible (see section 3.3).

3.3. Inter-annotator agreement

Any evaluation of an automatic error detection system, including that described in this work, should be performed with an understanding of the difficulty of the error-detection task for human listeners. To obtain a clearer picture of this task, we therefore conducted an analysis of the inter-annotator agreement observed in the annotations collected. The level of (dis)agreement among human annotators may indicate the difficulty of the task, which may in turn influence the standards by which an automatic system should be judged.

For 268 of the 668 utterances annotated, i.e. approximately 40% of the dataset, annotators were unanimous in their label assignments; for the other 400 utterances (60%), at least one annotator chose a different label than the other(s) who annotated the same utterance. Agreement in label assignments was calculated for each pair of annotators who overlapped, i.e. labeled any of the same tokens, quantified in terms of percentage agreement (the number of tokens to which the two annotators assigned the same label, divided by the total number of tokens they both annotated), and Cohen's Kappa (κ) statistic [14]. As an overall measure of inter-annotator agreement for the entire annotated dataset, we compute the minimum, median, mean, and maximum values over all pairwise comparisons (see table 2).

Mean and median percentage agreement values near 55% indicate that annotators seem to agree about the accuracy of lexical stress realizations just slightly more than they disagree, and the mean and median κ values near 0.25 characterize the overall agreement as "fair" in the Landis and Koch schema [15]. However, the minimum and maximum κ values reveal that agreement between different pairs of annotators ranges from "poor" to "substantial" [15], as also reflected in the correspondingly large gap between the minimum and maximum percentage agreement observed. On the whole, then, it appears that inter-annotator agreement in this error annotation task is relatively low, though there seems to be considerable variation between individual annotators. Finer-grained analysis did not reveal any explanation for this variation based on annotator L1 or expertise level. This low agreement may simply signal that (some of) the particular annotators participating in this study are not very reliable in their judgments of lexical stress accuracy, but it may also indicate that assessing L1 French speakers' realizations of German lexical stress is a difficult task, even for humans.

3.4. Error distribution

From the set of labels assigned to each word utterance by different annotators, a single "gold-standard" label for each utterance ultimately had to be chosen, as a representation of the ground truth with which to train and evaluate the automatic error classifier(s). In some cases, assigning a gold-standard label was trivial (e.g. when all or a majority of annotators agreed), but in others a choice had to be made between competing candidates. Label choice prioritized experts' judgments, favored confident judgments ([correct],[incorrect]) over [none], and gave learners the benefit of the doubt when annotators disagreed as to whether the utterance was [correct] or [incorrect].

Figure 1 illustrates the overall distribution of lexical stress errors in the annotated dataset with reference to the gold-standard labels thus determined. Most utterances were labeled [correct] (426 utterances, i.e. 63.8% of the 668 labeled utterances); in other words, almost two-thirds of the time, learners clearly stressed the correct (initial) syllable. However, learners also seemed to make mistakes regularly, with 29.6% (198) of their utterances labeled [incorrect] and another 5.2% (35) labeled [none]. (Eight, or 1.2%, of the utterances were labeled [bad_nsylls], and only one [bad_audio].) If we consider both [incorrect] and [none] utterances as types of lexical stress errors, then errors were observed in just over one-third of utterances. This considerable proportion of errors seems to confirm the expectation (mentioned in section 2) that French learners of German frequently make lexical stress errors.

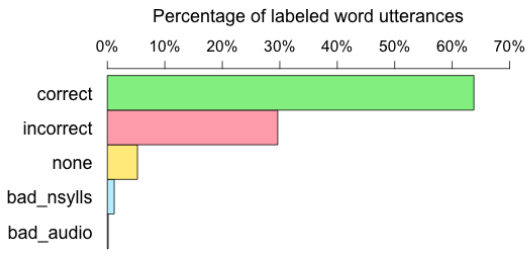

Figure 1: Distribution of gold-standard labels.

4. Method

By way of a preliminary investigation of the feasibility of classification-based identification of lexical stress errors in L2 German, a series of classification experiments were conducted in an effort to determine how accurately lexical stress productions can be automatically classified, and which features are most useful for this classification.

The WEKA machine learning toolkit[2] was used to train and evaluate simple Classification And Regression Tree (CART) classifiers for these experiments. Many other classification algorithms are implemented in WEKA, some of which could conceivably offer better performance, but CARTs were chosen for their simple training process and their ease of interpretation.

Using the features and training datasets described below (sections 4.1 and 4.2), CARTs were trained to classify utterances into one of the five categories described in section 3.2, with classification accuracy assessed via cross-validation.

4.1. Feature sets

To represent the lexical stress prosody of an utterance, the automatically-determined word, syllable, and phone segmentations were used to isolate relevant segments of the speech signal, and extract features related to duration, fundamental frequency (F0), and intensity.

Research on the phonetic realization of lexical stress has often indicated that duration may be the most important, if not the only, acoustic correlate of this phenomenon in German, with the duration of stressed syllables being relatively long in comparison with unstressed syllables [6]. Therefore, features representing duration were computed from the durations of relevant segments in the phone- and syllable-level segmentations. Following Bonneau and Colotte [8], we take into account the durations of entire syllables, as well as of their nuclei (vowels or syllabic consonants such as /ŋ/), as described in table 3a.

After duration, the next best acoustic correlate of lexical stress appears to be F0 [6], so F0 features were also computed (see table 3b). The F0 contour of a given utterance is estimated using the pitch detection functionality of the speech-processing program JSnoori[3] [16], which uses a spectral comb method to compute pitch points from spectra extracted from the relevant signal segment via Fast Fourier Transform (using Hamming windows 32 milliseconds long, offset by 8 ms). Pitch was computed in Hertz, then converted to semitones before features were computed. Features only take into account non-zero points, i.e. those corresponding to voiced segments. Though work on assessing L2 English stress has often made the assumption that stressed syllables should have higher F0 than unstressed ones (e.g. [8]), in German stressed syllables may also have a lower F0 than other syllable(s) in the word [1, p. 267]. Therefore, as table 3b shows, our features capture not only the maximum F0 in each syllable (nucleus), but also the minimum and range.

Past research indicates that a signal's intensity also reflects lexical stress patterns, though to a lesser extent than duration or F0 [1, 6]. Therefore, intensity contours of syllables and their nuclei were also calculated, again using JSnoori, and used to compute features taking into account the mean and maximum intensity in the relevant segments, as shown in table 3c.

Using different combinations of these feature sets, a series of experiments was conducted to determine which features give the best accuracy in the error-classification task. Establishing which features perform best not only enables the creation of the most accurate classification-based error diagnosis system possible, but may also clarify whether and how strongly these acoustic properties correspond with perceived lexical stress errors in L2 German. In addition to the prosodic features (table 3), features representing the uttered word and characteristics of the speaker (see table 4) were also included in the experiments, to ascertain whether such features could improve performance. The results of these experiments are presented in section 5.

[2] www.cs.waikato.ac.nz/ml/weka/

[3] jsnoori.loria.fr

4.2. Datasets for training and testing

As mentioned above, the labeled dataset described in section 3 was used as training and test data. In addition to this L2 data, L1 utterances of the selected word types from the the IFCASL-GG corpus were also included in training data, each having been automatically labeled as [correct] with the assumption that L1 speakers always realize stress correctly.

To evaluate the performance of the features described in

Table 3: Prosodic features. S0 refers to the word's first syllable, S1 to the second syllable; similarly, V0 and V1 refer to the nucleus (vowel) of the first and second syllable, respectively.

(a) Duration (DUR) feature set

Feature	Description
REL-S-DUR	Duration of S1/duration of S0
REL-V-DUR	Duration of V1/duration of V0

(b) Fundamental frequency (F0) feature set

Feature	Description
REL-S-F0-MEAN	Mean F0 in S1/ mean F0 in S0
REL-S-F0-MAX	Maximum F0 in S1/ max. F0 in S0
REL-S-F0-MIN	Minimum F0 in S1/ min. F0 in S0
REL-S-F0-RANGE	F0 Range (max. F0−min. F0) in S1/ F0 range in S0
REL-V-F0-MEAN	Mean F0 in V1/mean F0 in V0
REL-V-F0-MAX	Max. F0 in V1/max. F0 in V0
REL-V-F0-MIN	Min. F0 in V1/min. F0 in V0
REL-V-F0-RANGE	F0 Range in V1/F0 range in V0

(c) Intensity (INT) feature set

Feature	Description
REL-S-INT-MEAN	Mean intensity in S1/S0 mean int.
REL-S-INT-MAX	Maximum int. in S1/S0 max. int.
REL-V-INT-MEAN	Mean int. in V1/mean int. in V0
REL-V-INT-MAX	Max. int. in S1/max. int. in S0

Table 4: Speaker/word features

Feature	Description
WD	Word type uttered (e.g. *Tatort*; see table 1)
LV	Speaker's L2 German skill level (A2|B1|B2|C1)
AG	Speaker's age/gender (Girl|Boy|Woman|Man)

section 4.1, 10-fold cross-validation was performed on the entire set of available training data. To create each of the 10 folds, one-tenth of the L2 utterances were randomly selected to be held out as the test data, and the corresponding training set consisted of the remaining nine-tenths of the L2 utterances combined with the entire set of L1 utterances. Overall classification accuracy was computed by averaging the results over each of these 10 folds (see section 5).

To evaluate performance on unseen speakers, utterances from each of the 56 L2 speakers were held out in turn as testing data for a classifier trained on the L1 utterances as well as those of the other 55 L2 speakers, with overall accuracy computed by averaging the results over the 56 folds (see section 5).

4.3. Evaluation metrics

Classifier performance is quantified in terms of percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels. For the [correct] class, the following measures are also reported:

- Precision (P): number of utterances correctly classified as [correct] / total no. classified as [correct]
- Recall (R): no. correctly classified as [correct] / total no. of [correct] utterances in the gold-standard dataset
- F-measure (F_1): harmonic mean of P and R (where both are weighted equally): $F_1 = 2PR/(P+R)$
- F_2-measure: similar to F_1, but with R given twice as much weight as P: $F_2 = (1+2^2)\cdot PR/(2^2\cdot P+R)$

These are reported to account for the fact that in the intended application of CAPT, telling a student that they have made a mistake when in fact they have not can be more damaging to their motivation and willingness to continue learning with the system than telling them that they have stressed a word correctly when in fact they have made a mistake [4]. Therefore, [correct] R should be as close to 1 as possible, while still maintaining a balance with P such that the system does not trivially classify all utterances as [correct], which would render it useless. To keep this in perspective, the results in this section report both the commonly used F_1 measure, which weights P and R evenly, as well as F_2, which prioritizes R over P.

5. Results

Table 5 lists the results of experiments with the prosodic features described in table 3. As seen in the top rows of table 5, the results obtained using features representing each of the three acoustic correlates of lexical stress (duration, F0, and intensity) confirm that duration features seem to be the best predictor of lexical stress errors. In fact, the perfect (1.00) R values and κ at or near 0 for F0 and INT seem to indicate that these feature sets do not enable the system to discriminate between error classes at all, resulting in classifiers that are useless for CAPT insofar as they simply classify all utterances as [correct].

However, as the lower rows of table 5 show, better performance was obtained with classifiers trained on a combination of these features than using each set in isolation. Combining DUR and F0 gave the best overall performance using only prosodic features: 69.77% accuracy, κ = 0.29, and [correct] F_1 = 0.8.

As seen in table 6, even higher accuracy was obtained by combining prosodic features with the features representing speaker and word characteristics (see table 4). Information about the word type of the utterance (WD) and the L2 German proficiency of the speaker (LV) seemed to be most helpful, while including the speaker's age/gender (AG) appeared to have a negative, if any, impact on performance. Adding WD and LV to the best-performing prosodic features (DUR+F0) improved performance slightly; interestingly, however, the overall best performance on this dataset was achieved by combining WD and LV with the entire set of prosodic features (DUR+F0+INT), yielding average accuracy of 71.87%, κ of 0.34, and [correct]

Table 5: Results of experiments with prosodic features.The best values achieved for each metric are displayed in **bold**.

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR	66.78	0.19	0.69	0.91	0.79	0.86
F0	64.37	0.02	0.64	**1.00**	0.78	**0.90**
INT	63.77	0.00	0.64	**1.00**	0.78	**0.90**
INT+F0	64.52	0.04	0.65	0.98	0.78	0.89
DUR+INT	67.68	0.25	0.71	0.89	0.79	0.85
DUR+F0	**69.77**	**0.29**	**0.72**	0.91	**0.80**	0.86
DUR+F0+INT	67.52	0.25	0.71	0.89	0.79	0.85

Table 6: Results of experiments with speaker and word features. Best values achieved for each metric are displayed in **bold**.

(a) In combination with DUR+F0 feature set

Feature set (+DUR+F0)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WD	70.52	0.30	**0.72**	0.92	**0.81**	**0.87**
LV	68.72	0.27	0.71	0.91	0.79	0.86
AG	68.26	0.22	0.69	**0.94**	0.80	0.88
LV+AG	69.77	0.29	**0.72**	0.91	0.80	0.86
WD+AG	68.86	0.27	0.71	0.91	0.80	0.86
WD+LV	**70.65**	**0.31**	**0.72**	0.92	**0.81**	**0.87**
WD+LV+AG	68.41	0.26	0.71	0.91	0.79	0.86

(b) In combination with DUR+F0+INT feature set

Feature set (+DUR+F0+INT)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WD	68.41	0.28	0.72	0.88	0.79	0.84
LV	70.07	0.29	0.71	**0.92**	0.80	**0.87**
AG	66.93	0.24	0.71	0.88	0.78	0.84
LV+AG	68.57	0.27	0.72	0.89	0.79	0.85
WD+AG	68.87	0.30	**0.73**	0.87	0.79	0.83
WD+LV	**71.87**	**0.34**	**0.73**	**0.92**	**0.81**	**0.87**
WD+LV+AG	70.52	0.31	0.72	0.91	0.80	0.86

Table 7: Best results of experiments with unseen speakers.

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR+F0	69.16	0.19	0.68	0.90	0.74	0.85
DUR+F0+WD+LV	70.22	0.24	0.68	0.90	0.75	0.84

F_1 and F_2 measures of 0.81 and 0.87, respectively.

Though these results are the best of any of the experiments reported in this section, we would perhaps like to see better accuracy and F-measures, and higher than "fair" [15] agreement with the gold-standard labels, before placing such an error-diagnosis system in front of actual students. However, considering the relatively low agreement between humans tasked with the same type of error classification (see section 3.3), this accuracy does not seem so unimpressive. Indeed, the best average κ between the classifier output and gold-standard labels (0.34) exceeds the observed average human-human κ (0.23), and the best average percentage accuracy for that classifier (71.87%) is substantially higher than the average human-human percentage agreement (54.92%).

As would be expected, when classifying utterances from a speaker not seen in the training data (see section 4.2), accuracy drops slightly. As table 7 shows, using only DUR and F0 features resulted in 69.16% accuracy on unseen speakers, compared to 69.77% when speaker independence was not accounted for; however, a bigger drop in the κ value was observed, from 0.29 to 0.19. As with the randomly held-out data, better performance was obtained by adding WD and LV features, yielding 70.22% accuracy and $\kappa = 0.24$; again, this represents a slight drop from the 70.65% accuracy and $\kappa = 0.31$ observed using randomly split data. No improvements were observed when including INT or AG as additional features. Thus, the performance degradation when dealing with unknown speakers does not seem drastic, which is encouraging given that a CAPT system will of course need to classify speech from new users accurately. In future work it would be interesting to explore techniques for enabling improvements in accuracy as a learner continues to use the system (see Section 6.2).

6. Conclusions

Classification of lexical stress errors using machine learning algorithms is a novel approach to lexical stress error identification in German CAPT. This paper has explored how, and how effectively, classification-based diagnosis can be used to identify (in)correct realizations of lexical stress in the L2 German speech of L1 French speakers. The prosodic features found to be most useful for classification relate to duration and F0, unsurprising considering that past work has indicated these may be the closest acoustic correlates of lexical stress in German [1, 6]. Features representing the word type uttered and the L2 proficiency level of the speaker also seem valuable for error classification; combining these features with the three prosodic feature types yielded the overall highest accuracy (71.87% accuracy, κ = 0.34) attained on the L2 speech dataset (see section 3). Though these results leave room for improvement, they are encouraging given that agreement between the classifier's output and the gold-standard labels slightly exceeded the average agreement observed between human annotators asked to perform the same error classification task (see section 3.3). The findings reported here thus seem to confirm the utility of classification for diagnosing lexical stress errors in German CAPT, though further work is needed to achieve the level of performance necessary for real-world CAPT systems.

One logical direction for future work is the evaluation of other, more powerful machine learning algorithms than the simple CARTs used in this work; related work indicates that Maximum Entropy classifiers [11] and Neural Networks [12] may be promising. Consideration could also be given to additional features which may be related to lexical stress in German (e.g. those capturing vowel quality, phrase information, etc.). It may also be of interest to explore techniques for online, semi-supervised learning to improve accuracy as a learner uses the system over time, perhaps via an active learning approach soliciting teacher/expert judgments for certain utterances.

7. Acknowledgements

This work was partially supported by the IFCASL project (`ifcasl.org`), funded by the Deutsche Forschungsgemeinschaft and the Agence Nationale de la Recherche. We thank Bernd Möbius and the three anonymous reviewers for their helpful feedback on this work.

8. References

[1] A. Cutler, "Lexical Stress," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds., 2005, pp. 264–289.

[2] U. Hirschfeld, *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*, ser. Forum Phoneticum, 1994, vol. 57.

[3] M.-C. Michaux and J. Caspers, "The production of Dutch word stress by Francophone learners," in *Proc. of the Prosody-Discourse Interface Conference (IDP)*, 2013, pp. 89–94.

[4] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, 2002.

[5] U. Hirschfeld and J. Trouvain, "Teaching prosody in German as foreign language," in *Non-Native Prosody: Phonetic Description and Teaching Practice*, J. Trouvain and U. Gut, Eds. Walter de Gruyter, 2007, pp. 171–187.

[6] G. Dogil and B. Williams, "The phonetic manifestation of word stress," in *Word Prosodic Systems in the Languages of Europe*, H. van der Hulst, Ed. Walter de Gruyter, 1999, ch. 5, pp. 273–334.

[7] E. Dupoux, C. Pallier, N. Sebastian, and J. Mehler, "A Destressing 'Deafness' in French?" *Journal of Memory and Language*, vol. 36, no. 3, pp. 406–421, Apr. 1997.

[8] A. Bonneau and V. Colotte, "Automatic Feedback for L2 Prosody Learning," in *Speech and Language Technologies*, I. Ipsic, Ed. InTech, 2011.

[9] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 49, no. s1, pp. 285–310, 1999.

[10] J. Field, "Intelligibility and the Listener: The Role of Lexical Stress," *TESOL Quarterly*, vol. 39, no. 3, p. 399, Sep. 2005.

[11] Y.-J. Kim and M. C. Beutnagel, "Automatic assessment of American English lexical stress using machine learning algorithms," in *SLaTE*, 2011, pp. 93–96.

[12] 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.

[13] 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.

[14] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[15] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[16] J. Di Martino and Y. Laprie, "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal," in *EUROSPEECH*, 1999.