Automatic classification of lexical stress errors for German CAPT

Anjana Sofia Vakil

Department of Computational Linguistics & Phonetics
Saarland University, Saarbrücken, Germany
anjanav@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. Such 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 the first known exploration of the use of machine learning methods to classify non-native German lexical stress errors. Experiments using various prosodic **[TODO and other]** features yielded a **[TODO maximum]** classification accuracy of 71.87% on a manually-annotated corpus of German word utterances by native French speakers, which exceeded the observed inter-annotator agreement between humans asked to classify lexical stress errors in the same data. These results establish classification-based diagnosis of lexical stress errors as a viable approach for German CAPT. **[TODO Still got about 40 words - something about unseen word types?]**

Index Terms: computer-assisted pronunciation training, CAPT, word prosody, German **[TODO are these OK?]**

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, and has been found to have an impact on 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, 4].

To overcome this difficulty and improve their L2 word prosody, learners typically need to have their pronunciation errors pointed out and corrected by a language instructor; unfortunately, the lack of attention typically given to pronunciation in the foreign language classroom, along with other factors such as high student-to-teacher ratios, make this level of individualized attention not always feasible in a classroom setting [5, 6, 7]. 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 feedback on how to correct them, and thus to help learners achieve more intelligible pronunciation in the target language. However, while much research has gone into the creation and improvement of CAPT systems for English (see e.g. [8, 9]), relatively little work has been done on the development of CAPT systems for German, especially on those targeting errors in German prosody.

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 the creation of CAPT systems that address this error specifically, and summarizes some past work related to this topic. Section 3 describes the manual annotation of lexical stress errors in a small corpus of L2 German speech, carried out to create labeled training and test data for the classification experiments explained in section 4. Section 5 presents and analyzes the results of these experiments. Finally, section 6 offers some concluding remarks and outlines 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], i.e. how a syllable is given a more prominent role such that this syllable is perceived as "standing out" [10]. This perceived prominence of a syllable is a function not merely of the segmental characteristics of the uttered syllable, i.e. the speech sounds it contains, but rather of its (relative) suprasegmental properties, namely:

- duration, which equates on the perceptual level to length;
- fundamental frequency (F0), which corresponds to perceived pitch; and
- energy, which perceptually equates to loudness.

In variable-stress languages, such as German and English, the location of lexical stress in a word is not always predictable, and therefore 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 drive around) from *umFAHRen* (to run over with a car) in German. Furthermore, in German, misplaced stress can disrupt understanding even in cases where there is no stress-based minimal pair [2]. However, in fixed-stress languages, stress is completely predictable, as it always falls on a certain position in the word (e.g. the final syllable), making the lexical stress pattern less crucial to the knowledge of a word than in variable-stress languages. Furthermore, in fixed-stress languages there may be a weaker distinction between stressed and unstressed syllables. While French has often been categorized as a fixed-stress language, given that word-final syllables are given prominence

when a French 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, 4]. Regardless, French has no contrastive word-level stress [4, p. 89], and in this respect differs considerably from German.

This difference between the languages leads us to expect French learners of German to have difficulties with both perception and production of lexical stress prosody. Although little research has been done on the nature of lexical stress errors for this particular L1-L2 pair, Hirschfeld and Trouvain [7] report that such errors are commonly observed in German spoken by French natives. Research on French speakers' perception of Spanish, another contrastive-stress language, has revealed that these speakers seem to be "deaf" to lexical stress, i.e. seem to have significant and lasting difficulty perceiving and remembering stress contrasts [3]. With respect to production, studies of L2 Dutch have shown that French speakers, especially beginners, make systematic errors with lexical stress, exhibiting a tendency to stress the final syllable of Dutch words even when stress should be placed on the initial or medial syllable [11, 4]. Similar findings have also been reported for French learners of English [12]. The high (anticipated) frequency of lexical stress errors in the speech of this L1-L2 group is thus one motivating factor for the creation of CAPT systems to help learners identify and correct such errors.

Another motivation behind this work's focus on lexical stress errors is the high impact such errors may have on the intelligibility of L2 German speech. Intelligibility, as opposed to lack of a foreign accent, is generally considered to be the most important goal of pronunciation training [13, 5, 6, 14, 9]. The exact definition of *intelligibility* is a topic of debate, but here we will follow Munro and Derwing [13, p. 289] in understanding it broadly as "the extent to which a speakers message is actually understood by a listener." Generally speaking, prosodic errors have often been found to have a larger impact on the perceived intelligibility of L2 speakers than segmental errors [6, 9], and several studies have found lexical stress errors to have a particularly strong impact on intelligibility in free-stress languages like English and Dutch [1, 14] Though relatively little research has been done on how various pronunciation errors affect intelligibility in L2 German specifically, some studies suggest that lexical stress errors may hinder intelligibility of L2 German speech more than other types of errors [2, 7]. Stress errors may also affect perception of segmental errors in the L2 learners speech; for example, segmental errors occurring in stressed syllables may be more noticeable than those in unstressed syllables [1, 11].

Though the frequency and impact of lexical stress errors in the speech of French learners of German thus constitute strong reasons to develop CAPT tools to treat such errors, in order for such systems to be viable, the feasibility of reliable automatic detection of this type of error must be demonstrated. **[TODO Make rest of this par about how comparison-based diagnosis is usually used, then start new par with following sentence?]** To our knowledge, no work has been reported on automatic classification-based diagnosis of lexical stress 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 words. Kim and Beutnagel [15] experimented with various classifiers to identify stress patterns in high-quality recordings of 3- and 4-syllable English words, reporting accuracy in the 80-90% range; in pilot experiments with low-quality recordings, however, the authors report lower accuracy: 70-80% on L1 speech and 50-60% on utterances by L2 speakers. Similarly, Shahin et al. [16] trained Neural Networks to classify stress patterns in bisyllabic words uttered by L1 English children, and reported classification accuracy over 90% for some stress patterns; though this work was conducted with a view to treating childhood L1 dysprosody, its relevance to our intended application of L2 CAPT is nonetheless clear. Building on these related investigations, this paper seeks to further explore the viability of automatic classification-based detection of lexical stress errors, with a particular focus on those made by French speakers of German. To this end, a small corpus of learner speech was manually annotated for lexical stress errors, as described in section 3. Using the resulting labeled L2 data, in addition to data from L1 German speakers, a series of supervised machine learners were trained using a variety of representations of the prosodic and other features of each word utterance (see section 4.1, and these classifiers were evaluated with reference to the manually-produced labels of held-out test data (see section 4.2). Section 5 presents and analyzes the results of these evaluations.

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 by native and non-native German speakers with varying levels of phonetics/phonology expertise. This section describes the data selected for annotation (section 3.1) and the method by which lexical stress realizations in this data were annotated (section 3.2), and presents an analysis of the observed inter-annotator agreement (section 3.3) and distribution of errors (section 3.4) in the annotated dataset.

3.1. The IFCASL corpus of learner speech

The learner speech data used in this work has been excerpted from the IFCASL corpus [17], 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. This is the first known corpus of L2 speech in both directions of the French-German language pair, and is thus an invaluable resource for research on L2 pronunciation errors in these languages.

This corpus contains recordings of approximately 50 L1 speakers of each language reading carefully constructed 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 self-reported L2 proficiency levels are also represented in the corpus: the recorded adults span CEFR[1] levels A2 (beginner) through C1 (advanced), the children levels A2 (beginner) and B1 (low intermediate).

While L2 French speech is thus also captured in the IFCASL corpus, the annotation effort described here focuses exclusively on the German-language subset of the corpus. Only

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

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

Word	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
Frhling	/ˈfryː.lɪŋ/	noun	spring (season)	56
halten	/ˈhal.tn̩/	verb	to hold	56
manche	/ˈman.çə/	pronoun	some	56
Mrder	/ˈmœʁ.dɐ/	noun	murderer	56
Pollen	/ˈpɔ.lən/	noun	pollen	55
Ringen	/ˈʁɪŋ.ən/	noun	rings	55
Tatort	/ˈtaːt.ʔɔʁt/	noun	crime scene	56
tragen	/ˈtʁaː.gn̩/	verb	to wear	55
Tschechen	/ˈtʃɛ.çn̩/	noun	Czechs	56

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.

The subset of IFCASL-FG selected for manual error annotation, (henceforth simply 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, given the phenomenon of final lengthening in French (see section 2).

In addition to the recordings themselves, the IFCASL corpus contains phone- and word-level segmentations of each utterance, produced automatically by forced alignment with the corresponding text prompts [17]. Although the corpus also contains manual corrections of these segmentations, the work reported here relies exclusively on the automatically-generated segmentations to more accurately represent the conditions of a real-world CAPT system, which would not have recourse to manual verification of the phone or word boundaries identified by the aligner. As the IFCASL corpus does not include syllable-level segmentations, boundaries between syllables were determined automatically from the phone- and word-level segmentations of each utterance.

Using the automatic segmentations, tokens (utterances) of each selected word type were extracted from the recorded segmentations automatically using Praat[2]; counts of the available tokens for each word type are listed in table 1. The dataset annotated for lexical stress errors comprises 668 word tokens in total. Five tokens had to be excluded from the dataset, as sentence-level disfluencies (e.g. false starts or repetitions of phrases) prevented accurate automatic extraction of the word utterance; a fully-fledged CAPT system would need to deal with such disfluencies automatically, e.g. with a pre-processing step which detects disfluencies and prompts the learner to re-record their utterance if needed (see e.g. [18]).

[2] praat.org

3.2. Annotation method

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

- [correct]: the speaker clearly stressed the correct (initial) syllable
- [incorrect]: the speaker clearly stressed the incorrect (final) syllable
- [none]: the speaker did not clearly stress either syllable, or the annotator was unable to determine which syllable was stressed
- [bad_nsylls]: the speaker pronounced an incorrect number of syllables (e.g. inserted an extra syllable), making it impossible to judge if stress was realized correctly
- [bad_audio]: a problem with the audio file (e.g. noise or inaccurate segmentation) interfered with the annotator's ability to judge the stress realization

Annotation was performed using a graphical tool scripted in Praat. This tool displayed the given word's text, and allowed the annotator to listen to the given word utterance and the sentence utterance from which it was extracted as many times as they wished. Once they had reached a judgment about the lexical stress realization of the utterance, the annotator had to click 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.

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 to maximize 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 as described in the previous section. If human annotators often disagree about whether a given L2 utterance contains a lexical stress error, this may indicate that the task is a difficult one, thus encouraging a more

Table 2: Overall pairwise agreement between annotators

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

lenient evaluation of an automatic error-detection system. However, if humans are generally in strong agreement, this may reflect a lower level of difficulty, and give reason to judge the performance of an automatic system by a higher standard.

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. To make sense of these differences, agreement in label assignments was calculated for each pair of annotators who overlapped, i.e. labeled any of the same tokens. Pairwise agreement was quantified in terms of percentage agreement (i.e. 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 [19]. To obtain an overall measure of inter-annotator agreement for the entire annotated dataset, the agreement between each pair of overlapping annotators was calculated, and the minimum, median, mean, and maximum values over all pairwise comparisons were computed; these values are given in table 2.

This simple analysis reveals a few interesting observations. First, the 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 characterizes the overall agreement as "fair" in the Landis and Koch schema [20]. However, the minimum and maximum κ values reveal that agreement between different pairs of annotators ranges from "poor" to "substantial" [20], 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. This 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 diagnosing errors in L1 French speakers' realizations of lexical stress in German 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. However, in other cases a choice had to be made between competing candidate labels. 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].

To shed further light on the error classification task, the

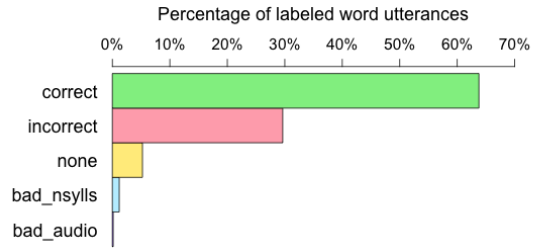

Figure 1: Overall distribution of lexical stress errors in the annotated data

overall distribution of lexical stress errors in the annotated dataset was analyzed with reference to the gold-standard labels thus determined. As illustrated in fig. 1, most of learners' lexical stress realizations were judged to be 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 in the uttered word. However, learners also seemed to make mistakes regularly, with 29.6% (198) of their word utterances labeled [incorrect] and another 5.2% (35) labeled [none]. (Eight, or 1.2%, of the word 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.

4. Method

As mentioned in section 2, the classification-based approach to identifying lexical stress errors has not been sufficiently explored in CAPT research, especially research on CAPT for German. By way of a preliminary investigation of the feasibility of this type of error diagnosis, a series of classification experiments were conducted in an effort to determine: Therefore, we conducted a series of experiments to investigate:

- how accurately lexical stress errors can be automatically classified,
- which of the features discussed in section 4.1 are most useful for this classification, and
- whether classification can enable accurate error diagnosis for words not seen in the training data.

The WEKA machine learning toolkit[3] was used to train and evaluate classifiers for these experiments. In this work, only simple Classification And Regression Tree (CART) classifiers were used. Many other classification algorithms are implemented in WEKA, some of which could conceivably offer better performance; here, CARTs were chosen for their simple training process and their ease of interpretation by humans.

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. In practice, however, the classifiers only assign the labels [correct] and [incorrect], apparently neglecting the others due to their comparatively low frequency in the data. Overall classification accuracy was assessed by holding out portions of the an-

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

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 (DR) feature set

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

(b) Fundamental frequency (F0) feature set

Feature name	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) Energy (EN) feature set

Feature name	Description
REL-S-EN-MEAN	Mean energy in S1/S0 mean en.
REL-S-EN-MAX	Maximum en. in S1/S0 max. en.
REL-V-EN-MEAN	Mean en. in V1/mean en. in V0
REL-V-EN-MAX	Max. en. in S1/max. en. in S0

notated data for testing, and performing n-fold cross-validation (see section 4.2).

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 the three acoustic properties mentioned in section 2 above: duration, fundamental frequency (F0), and energy. To account for inter-speaker variability, e.g. the fact that some speakers may have a faster speech rate or higher F0 than others, relative rather than absolute features were used. This section explains how these prosodic features, listed in table 3, were selected and computed.

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 [10]. Therefore, features representing duration were computed, simply by noting the relative durations of relevant segments in the phone- and syllable-level segmentations. Following Bonneau and Colotte [12], we take into account the durations of entire syllables, as well as of their nuclei (vowels or syllabic consonants such as /n̩/), as described in table 3a.

After duration, the next best acoustic correlate of lexical stress appears to be F0 [10], 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[4] [21], which uses a spectral comb method to compute pitch points from each FFT spectrum extracted from the relevant signal segment; here spectra were extracted using Hamming windows 32-milliseconds long, offset by 8 ms. Points are computed in Hz, then converted to semitones before features are 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. [12]), 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 (maximum−minimum).

Table 4: Speaker/word features

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

Past research indicates that a signal's energy (intensity) also reflects lexical stress patterns, though to a lesser extent than duration or F0 [10, 1]. Therefore, energy contours of syllables and their nuclei were also computed, again using JSnoori, which calculates the total amount of energy at frequencies from 0 to 8000 Hz in FFT spectra extracted from the signal (using Hamming windows 20 ms. long offset by 4 ms). Energy values below 60 decibels are not counted toward the total, as these are assumed to correspond to ambient or non-speech noise. Using these contours, we compute features taking into account the mean and maximum energy 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 word type uttered and characteristics of the speaker (see table 4) were also included in the experiments, to ascertain whether including such features could improve performance. The results of these feature experiments are presented in section 5.1.

4.2. Datasets for training and testing

As mentioned above, CART classifiers were trained and tested using the labeled dataset described in section 3. 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] based on the assumption that native speakers always realize lexical stress correctly.

[4] jsnoori.loria.fr

To evaluate the performance of the features described in section 4.1, 10-fold cross-validation was performed on the entire set of available training data. The current objective being the classification of L2 speech, including L1 utterances in the test data was not appropriate; instead, 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 5.1).

[TODO Compared to comparison-based diagnosis], a classification-based approach to error diagnosis is attractive in that it can theoretically enable the creation of CAPT exercises for new word types without requiring additional recordings of L1 utterances of these words. To assess whether this extension to new words is really feasible, we also evaluated classification performance on words not seen in the training data. This was accomplished by dividing the annotated dataset into 12 subsets, one for word type (see table 1), and performing a 12-fold cross-validation. For each fold, the test dataset consisted of the L2 utterances of a single held-out word type, while the training dataset comprised the L1 and L2 utterances of the other 11 words.

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_1 measure: harmonic mean of P and R (where both are weighted equally), i.e. $F_1 = 2PR/(P+R)$
- F_2 measure: similar to F_1 measure, but with R given twice as much weight as P:

$$F_2 = (1+2^2) \cdot PR/(2^2 \cdot P + R) = 5PR/(4P+R)$$

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 [5]. 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

This section describes the results of the classification experiments conducted by training CART classifiers using various configurations of the feature sets and train/test data splits described in section 4.

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
DR	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**
EN	63.77	0.00	0.64	**1.00**	0.78	**0.90**
EN+F0	64.52	0.04	0.65	0.98	0.78	0.89
DR+EN	67.68	0.25	0.71	0.89	0.79	0.85
DR+F0	**69.77**	**0.29**	**0.72**	0.91	**0.80**	0.86
DR+F0+EN	67.52	0.25	0.71	0.89	0.79	0.85

5.1. Feature performance

Table 5 lists the results of experiments with the prosodic features described in table 3. As seen in the first three rows of table 5, the results obtained using features representing each of the three acoustic correlates of lexical stress, duration, F0, and intensity (energy), 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 energy features 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. The pairing of duration and F0 features, with energy features omitted, resulted in 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)[5] seemed to be most helpful, while including the speaker's age/gender category (AG) appeared to have a negative, if any, impact on performance. Adding WD and LV to the best-performing prosodic features (DR+F0) improved classification accuracy slightly across all metrics; interestingly, however, the overall best performance on this dataset was achieved by combining WD and LV with the entire set of prosodic features (DR+F0+EN), yielding average accuracy of 71.87%, κ of 0.34, and [correct] F_1 and F_2 measures of 0.81 and 0.87, respectively.

Though these statistics 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" agreement [20] 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)

[5] **[TODO does this fit?]** Though including the LV feature would be nonsensical if the intended application of the classifier were assessment of a learner's proficiency level, the goal here is not assessment but training, so in this case it is reasonable to allow the system to take LV into account.

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

(a) In combination with DR+F0 feature set

Feature set (+DR+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 DR+F0+EN feature set

Feature set (+DR+F0+EN)	% 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

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%).

5.2. Performance on unseen words

Table 7 presents the results of experiments with unseen words (see section 4.2), averaged over each of the 12 held-out word types. Interestingly, the highest accuracy (66.85%) was achieved using classifiers trained only on the best-performing prosodic features (DR+F0), excluding the speaker-related features (LV and AG). However, classifiers trained on all three prosodic feature sets (DR+F0+EN) yielded the best agreement with the gold-standard labels (κ=0.19), while the best F-measures (F_1=0.77, F_2=0.85) were observed using classifiers trained using all available features (DR+F0+EN+LV+AG; WD was of course not included in these experiments).

Table 7: Results of experiments with unseen words. The best values achieved for each metric are displayed in **bold**.

Feature set	% acc.	κ	[correct] class P	R	F_1	F_2
DR+F0	**66.85**	0.17	0.69	0.88	**0.77**	0.84
+LV	65.51	0.16	0.69	0.89	**0.77**	0.84
+AG	65.05	0.16	0.69	0.88	0.76	0.84
DR+F0+EN	65.66	**0.19**	**0.70**	0.85	0.76	0.82
+LV	64.16	0.11	0.67	0.88	0.75	0.83
+AG	64.31	0.12	0.68	**0.90**	**0.77**	**0.85**

The drop in performance compared with the best results obtained on seen words (see table 6b) seems to indicate that extending classification to words for which L1 utterances are not available for training is not a trivial matter. The difference in κ is perhaps most striking, constituting a drop from "fair" to "slight" agreement with the gold standard [20], as well as a fall below the average inter-annotator κ (0.23). However, the percentage accuracy still exceeds the average percentage agreement between human annotators (54.92%), and the observed decreases in percentage accuracy and F-measures **[TODO with respect to seen words]** do not seem drastic. This is encouraging for the prospect of classifying errors in novel word types; if overall classification performance can be improved by other means (e.g. more powerful machine learning algorithms, additional features), classifying errors in unseen words **[TODO may yet be feasible]**.

6. Conclusions and future work

Classification of lexical stress errors using machine learning algorithms is a relatively 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 [10, 1]. Features representing the word type uttered and the L2 proficiency level of the speaker also seemed 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). Somewhat lower classification accuracy was observed **[TODO with utterances of word types]** not represented in the training data, although this accuracy still remained comparable to the average percentage agreement between humans. The findings of the work reported here thus seem to **[TODO establish]** the utility of classification-based diagnosis for 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 [15] and Neural Networks [16] may be promising. Consideration could also be given to additional features which may be related to lexical stress in German (e.g. spectral features capturing aspects of vowel quality). It may also be of interest to explore ensemble approaches to error classification, in which labels assigned by different classifiers trained only on subsets of the available features are compared to arrive at a final decision.

7. 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] E. Dupoux, N. Sebastián-Gallés, E. Navarette, and S. Peperkamp, "Persistent stress 'deafness': The case of French learners of Spanish," *Cognition*, vol. 106, pp. 682–706, 2008.

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

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

[6] T. M. Derwing and M. J. Munro, "Second Language Accent and Pronunciation Teaching: A Research-Based Approach," *TESOL Quarterly*, vol. 39, no. 3, pp. pp. 379–397, 2005.

[7] 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.

[8] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[9] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 1–8.

[10] 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. Berlin: Walter de Gruyter, 1999, ch. 5, pp. 273–334.

[11] M.-C. Michaux, "Exploring the production and perception of word stress by French-speaking learners of Dutch," in *Workshop on Crosslinguistic Influence in Non-Native Language Acquisition*, 2012.

[12] A. Bonneau and V. Colotte, "Automatic Feedback for L2 Prosody Learning," in *Speech and Language Technologies*, I. Ipsic, Ed. InTech, 2011, no. 1977.

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

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

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

[16] M. Shahin, B. Ahmed, and K. Ballard, "A neural network based lexical stress pattern classifier," *Qatar Foundation Annual Research Forum Proceedings*, no. 2012, p. CSP22, Oct. 2012.

[17] 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.

[18] L. Orosanu, D. Jouvet, D. Fohr, I. Illina, and A. Bonneau, "Combining criteria for the detection of incorrect entries of non-native speech in the context of foreign language learning," in *SLT 2012 - 4th IEEE Workshop on Spoken Language Technology*, Dec. 2012.

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

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

[21] 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 *6th European Conference on Speech Communication & Technology (EUROSPEECH99)*, Budapest, Hungary, 1999, p. 4.