Summary of the M.Sc. thesis

Automatic diagnosis and feedback for lexical stress errors in non-native speech: Towards a CAPT system for French learners of German

May 14, 2015

1 Introduction

The prosodic realization of lexical stress, the phenomenon by which certain syllable(s) in a word are accentuated more than others, is an important feature of the German phonological system, but one that can pose a considerable challenge to students learning German as a foreign language (L2), especially students whose native language (L1) is French. This thesis investigates how Computer Assisted Pronunciation Training (CAPT) can help L1 French speakers improve their lexical stress prosody in German. It describes the manual annotation of lexical stress errors in a small learner speech corpus, investigates a variety of methods for automatically diagnosing such errors in learner word utterances, and explores how these diagnoses can be used to deliver diverse types of feedback to learners. Its most important contribution is a prototype CAPT tool called **de-stress**: the German (**de**) **S**ystem for **T**raining and **R**esearch on **E**rrors in **S**econd-language **S**tress. This tool integrates various diagnostic and feedback methods via an easy-to-use web interface, and is designed not only to provide students with automatic, individualized feedback on their lexical stress errors, but also to facilitate future research on the efficacy of the different diagnostic and feedback approaches explored in this thesis.

2 Lexical stress errors by French learners of German

In an effort to shed light on the nature of lexical stress errors in the speech of L1 French learners of German as L2, lexical stress realizations in utterances of bisyllabic, initial-stress words taken from the IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013) were annotated by native and non-native German annotators with various levels of phonetics/phonology expertise.Annotators were asked to use a graphical annotation tool to label each recorded word utterance as correctly or incorrectly realizing lexical stress (i.e. the speaker clearly stressed the correct or incorrect syllable), failing to clearly realize stress (i.e. the speaker did not seem to stress either syllable), or having technical or other problems which prevented the assessment of lexical stress.

Analysis of the labels assigned by different annotators to the same utterances revealed relatively low inter-annotator agreement; On average, pairs of annotators who labeled the same utterances agreed on **[TODO X]** percent of utterances, corresponding to an average Cohen's Kappa (Cohen, 1960) value of **[TODO X]**. Considerable variation was observed among individual annotators, which did not seem to be explained by differences in their L1 or level of phonetics/phonology expertise. **[TODO remove?]** There also seemed to be variability in inter-annotator agreement with respect to the different word types in the dataset, and further work is needed to discern the factors responsible for this observation.

The multiple, often conflicting, error annotations from different annotators were consolidated into a single gold-standard annotation for each utterance in the dataset, which served as the basis for an analysis of the frequency and type of errors produced by learners. Overall, approximately two-thirds **[TODO (X%)]** of learners' utterances were deemed to realize lexical stress correctly, confirming that French learners of German frequently make errors with respect to lexical stress. The observed frequency of such errors was considerably lower in the speech of advanced learners than that of beginners, and children (ages 15-16) seemed to make more errors than adult beginners; no substantial difference was observed between speakers of different genders. As in the case of inter-annotator agreement, considerable variation was observed in the frequency of errors in utterances of different

word types, though once again the factors underlying this variability are not immediately evident and should be investigated in future work.

The error annotation and analysis described in this chapter thus contribute considerably to our understanding of the difficulties L1 French speakers may have realizing lexical stress in German, and the fact that learners, especially beginners and children, seem to struggle with lexical stress production justifies the selection of lexical stress errors as the focus of this thesis project. Additionally, the analysis of inter-annotator agreement presented in this chapter, specifically the finding that the observed agreement was generally rather low, constitutes an important discovery with respect to the task of identifying such errors in learner speech: though further research is needed to determine why and under which conditions this is the case, it would seem that diagnosing lexical stress errors may be a challenging task for at least some L1 and L2 German speakers. If true, this has important implications for the development and evaluation of automatic error diagnosis systems, which is the subject of the following section.

3 Diagnosis of lexical stress errors

With the motivation of helping French learners of German correct these frequently-produced lexical stress errors, this thesis explores the methods by which such errors can be automatically diagnosed in learner speech. All diagnosis starts from a prosodic analysis of the learner's utterance in terms of duration, fundamental frequency (F0), and intensity, i.e. the three acoustic properties most strongly correlated with the prosodic realization of lexical stress (see e.g. Cutler, 2005; Dogil and Williams, 1999). This analysis requires reasonably accurate automatically-produced segmentations of the words, syllables, and phones of a given utterance; while forced alignment can theoretically be used to produce such segmentations on the fly, the current implementation of de-stress mocks the alignment step by using pre-existing automatic segmentations for the utterances from the IFCASL corpus (Fauth et al., 2014). Using these segmentations, features capturing the relative duration, F0, and intensity of each syllable in a given word utterance are automatically extracted using the speech processing functionality of the JSnoori software[1] (Di Martino and Laprie, 1999; Laprie, 1999).

[1] `jsnoori.loria.fr`

Using these features to represent a given learner's (L2) utterance as well as corresponding utterance(s) by L1 speakers, it is possible to assess the learner's utterance via one of two primary strategies: comparison or classification. Diagnosis by comparison is the method most commonly used in past CAPT systems and research (see e.g. Bonneau and Colotte, 2011; Delmonte, 2011; Eskenazi, 2009). In the simplest comparison-based approach, features of the relevant segments of a learner's utterance are compared to the analogous features in a single L1 (reference) utterance using JSnoori, and an error is diagnosed when the utterances differ considerably with respect to the relevant features. To reduce some of the risk of the diagnosis "over-fitting" to speaker- or utterance-dependent features of a single reference, de-stress can average the results of multiple one-on-one comparisons to produce a multiple-reference diagnosis; this is a relatively novel approach to comparison-based diagnosis. Another innovation of de-stress relates to the method of choosing the reference utterance(s) for a given learner utterance; though many existing CAPT systems, including JSnoori, require manual selection of reference utterances, previous research by Probst et al. (2002) suggests that using an intelligently-selected reference speaker can help learners improve their pronunciation. Therefore, de-stress also offers an automatic selection option in which the reference is selected by choosing the L1 speaker(s) whose voice most closely resembles that of the learner, in terms of F0 mean and range.

The second diagnosis strategy explored in this thesis, classification of errors using machine learning algorithms, is a more novel approach to lexical stress error identification in CAPT, in which a learner's utterance is compared to the more abstract model of L1 speech represented by a classifier trained on a large number of L1 utterances. The classification experiments conducted in this work constitute original contributions to the understanding of how, and how effectively, classification-based diagnosis can be used to identify (in)correct realizations of lexical stress in German. Experiments with different prosodic feature sets **[TODO (see ??)]** suggest that the features seemingly most useful for classification relate to the duration and F0 of the utterance(s), unsurprising considering that these have been shown to be most closely linked to lexical stress in German (Cutler, 2005; Dogil and Williams, 1999). Features capturing the word type of the utterance as well as the age, gender and proficiency level of the speaker were also

found to be quite valuable for error classification; combining these features with all three prosodic feature types resulted in the highest overall accuracy observed on this dataset **[TODO (70.65% accuracy, κ =0.31)]**. As the observed agreement between the classifier's labels and the gold standard thus slightly exceeded the overall inter-annotator agreement observed when humans were asked to perform this error diagnosis task (see section 2), these results seem encouraging. Unsurprisingly, slightly lower accuracy was observed when classifying utterances of word types or speakers not represented in the training data; however, the fact that accuracy on unseen words still remained comparable to the human inter-annotator agreement statistics seems to confirm the expectation that classification-based diagnosis may be a useful way to create CAPT systems which are are not limited to words/sentences for which recorded L1 utterances are available.

In the administrative interface to de-stress, a researcher (or instructor) can choose between these various approaches for diagnosing lexical stress errors (see **[TODO ??]**). This modularity and flexibility enables de-stress to facilitate much needed future research exploring which diagnostic methods are most useful to which learners in which situations.

4 Feedback on lexical stress errors

[TODO This chapter has presented] the diverse array of feedback methods offered by de-stress. The system can present learners with both explicit **[TODO (??)]** and implicit **[TODO (??)]** feedback on their lexical stress errors. The latter may be delivered visually **[TODO (??)]**, in the form of graphical or textual representations of prosody which may be easier to interpret than the more direct representations of speech signals more frequently seen in CAPT systems **[TODO (see ??)]**, or via the auditory channel **[TODO (??)]** in the form of original or prosodically modified utterances. Explicit feedback options include skill bars **[TODO (??)]** informing learners of the correctness of their pronunciation with regard to duration (timing), fundamental frequency (pitch), and intensity (loudness), as well as verbal feedback on this information via a series of error/success messages **[TODO (??)]**. An additional feedback option offers learners the opportunity to self-assess their pronunciation **[TODO (??)]**, in the hopes of encouraging them to develop a metacognitive understanding of their own learning process and

take autonomous control of their learning.

Thanks to its modular implementation of the various feedback types, in combination with a simple configuration interface that allows researchers and/or instructors to easily pick and choose from the available options **[TODO (??)]**, de-stress is therefore a valuable platform for future empirical investigations into the impacts of these feedback types on the acquisition of L2 word prosody by L1 French learners of German. As established in **[TODO ??]** and further discussed in the following chapter, much work remains to be done to determine which feedback types are most effective in which learning contexts; by presenting a tool to facilitate such research, ostensibly the first of its kind for this type of error and aimed at this group of learners, this thesis has thus made an important contribution towards a more detailed understanding of the relative efficacy of feedback types in CAPT.

5 Conclusion

[TODO This thesis has thus presented] original work taking steps towards the development of a comprehensive, intelligent CAPT system for French learners of German. Drawing from previous research on foreign-language pedagogy, phonetics/phonology, and speech technology, as well as original research on the frequency of lexical stress errors in learner speech and novel methods for diagnosing such errors, the de-stress system developed in this thesis project is the first known CAPT tool dedicated to helping L1 French speakers improve their realization of lexical stress in German, and to facilitating research on the use of automatic, individualized feedback to help learners correct such errors. **[TODO SOMETHING ABOUT INTEGRATION WITH ITS - The following section suggests some possible directions for future work on improving de-stress and using it to make further advancements in CAPT for German.]**

References

Bonneau, Anne and Vincent Colotte (2011). "Automatic Feedback for L2 Prosody Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on p. 4).

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 2).

Cutler, Anne (2005). "Lexical Stress". In: *The Handbook of Speech Perception*. Ed. by David B. Pisoni and Robert E. Remez, pp. 264–289 (cit. on pp. 3, 4).

Delmonte, Rodolfo (2011). "Exploring Speech Technologies for Language Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on p. 4).

Di Martino, Joseph and Yves Laprie (1999). "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal". In: *EUROSPEECH*. Budapest, Hungary, p. 4 (cit. on p. 3).

Dogil, Grzegorz and Briony Williams (1999). "The phonetic manifestation of word stress". In: *Word Prosodic Systems in the Languages of Europe*. Ed. by Harry van der Hulst. Berlin: Walter de Gruyter. Chap. 5, pp. 273–334 (cit. on pp. 3, 4).

Eskenazi, Maxine (2009). "An overview of spoken language technology for education". In: *Speech Communication* 51.10, pp. 832–844 (cit. on p. 4).

Fauth, Camille, Anne Bonneau, Frank Zimmerer, Jürgen Trouvain, Bistra Andreeva, Vincent Colotte, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella, and Bernd Möbius (2014). "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, pp. 1477–1482 (cit. on pp. 2, 3).

Laprie, Yves (1999). "Snorri, a software for speech sciences". In: *ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)*. London, UK, pp. 89–92 (cit. on p. 3).

Probst, Katharina, Yan Ke, and Maxine Eskenazi (2002). "Enhancing foreign language tutors - In search of the golden speaker". In: *Speech Communication* 37.3-4, pp. 161–173 (cit. on p. 4).

Trouvain, Jürgen, Yves Laprie, Bernd Möbius, Bistra Andreeva, Anne Bonneau, Vincent Colotte, Camille Fauth, Dominique Fohr, Denis Jouvet, Odile Mella, Jeanin Jügler, and Frank Zimmerer (2013). "Designing a bilingual speech corpus for French and German language learners". In: *Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages*. ii. Strasbourg, France, pp. 32–34 (cit. on p. 2).