Automatic diagnosis and feedback for lexical stress errors in non-native speech

Towards a CAPT system for French learners of German

Anjana Sofia Vakil

Thesis Proposal
M.Sc. Language Science and Technology

Supervisors:
Prof. Dr. Bernd Möbius
Dr. Jürgen Trouvain

Saarland University
Department of Computational Linguistics & Phonetics

5 September, 2014

1 Introduction

In the field of second-language education, pronunciation has traditionally been given less attention than other areas such as grammar or vocabulary (Derwing and Munro, 2005). One reason for this may be that pronunciation is best taught through one-on-one instruction, which is not often possible in the traditional classroom setting. Hence the attraction of Computer-Assisted Pronunciation Training (CAPT) systems, which have the potential to automatically provide highly individualized analysis of learner errors, and feedback on how to correct them and achieve more intelligible pronunciation in the target language (Witt, 2012).

For students with French as their first language (L1) who are learning German as a second language (L2), the sound system of the L2 can pose a variety of difficulties, one of the most important and interesting of which is the way in which certain syllables in German words are accentuated more than others, a phenomenon referred to as lexical stress. Learning to navigate German lexical stress is especially challenging for L1 French speakers, because this phenomenon is realized differently, or perhaps does not occur at all, in the French language.

With these motivations in mind, the proposed thesis project aims to advance German CAPT by creating a tool which will diagnose and offer feedback on lexical stress errors in the L2 German speech of L1 French speakers, in the hopes of ultimately helping these learners become more intelligible when speaking German.

1.1 Context: The IFCASL project

This work will be conducted in the context of the ongoing research project "Individualized Feedback in Computer-Assisted Spoken Language Learning (IFCASL)" at the University of Saarland (Saarbrücken, Germany) and LORIA (Nancy, France).

The ultimate goal of the IFCASL project is to take initial steps toward the development of a CAPT system targeting, on the one hand, native (L1) French speakers learning German as a foreign language (L2), and on the other, L1 German speakers learning French as their L2. To this end, a bidirectional learner speech corpus has been recorded, comprising phonetically diverse utterances in French and German spoken by both native speakers and non-native speakers with the other language as L1 (Fauth et al., 2014; Trouvain et al., 2013).

This thesis will focus exclusively on French L1 speakers learning German as L2. The German-language subset of the IFCASL corpus will be instrumental in training and testing the automatic diagnosis and feedback systems which this work aims to develop. Futhermore, those systems will be designed with a view to contributing to the overall set of software developed in the context of the IFCASL project, such that they will be as compatible as possible with the other tools developed and used by the IFCASL team.

1.2 Objectives

The main objective of this work is to investigate the automatic treatment of lexical stress errors in the context of a CAPT system for French learners of German. This includes, on the one hand, an examination of the ways in which lexical stress errors of the type made by French L1 speakers when speaking German as L2 can be reliably detected and measured automatically, and on the other, an exploration of the types of multimodal feedback on such errors that can be automatically delivered based on the aforementioned error detection.

The intented outcome of these investigations is a prototype CAPT toolwhich can diagnose lexical stress errors in different ways and present learners with different types of feedback on these errors, such that researchers can use this modular system to study the impact of various assessment and feedback types on learner outcomes, user engagement, and other factors impacting the success of a CAPT system.

Once more is known about which diagnosis/feedback types should be delivered to which learners in which situations, this tool could become a useful component to a fully-fledged CAPT system, in which learner models and other intelligent components automatically decide which modules of the tool to activate.

1.3 Proposal overview

The remainder of this proposal is structured as follows. Section 2 places this thesis in the context of existing research on CAPT, and motivates its specific focus on lexical stress errors. Section 3 outlines the techniques that will be explored for diagnosing lexical stress errors in learners' speech via automatic processing of acoustic correlates of these errors in a spoken utterance. Section 4 describes the multimodal feedback types the system will aim to deliver, and how these could be generated automatically from the analysis described in the previous section. Section 5 summarizes the proposal and the aims of the thesis.

2 Background and related work

2.1 Computer-Assisted Pronunciation Training

The educational value of speech technologies has been well demonstrated in recent decades (Eskenazi, 2009), with Computer-Assisted Pronunciation Training (CAPT) emerging as one important educational application for foreign-language education (FLE) (Neri et al., 2002; Delmonte, 2011; Witt, 2012). This section places CAPT in the context of foreign-language pronunciation instruction (sec. 2.1.1), and describes a few recent CAPT systems which incorporate training in prosody (sec. 2.1.2). **[TODO why prosody?]**

2.1.1 Pronunciation in foreign language education

In the foreign language classroom, less focus has traditionally been placed on pronunciation than other aspects of language education, such as grammar and vocabulary (Neri et al., 2002; Derwing and Munro, 2005). However, even when pronunciation is taught in the classroom, a number of factors may limit the effectiveness of that training (Neri et al., 2002; Derwing and Munro, 2005). First of all, partly thanks to a historical lack of communication between the fields of speech science and FLE, many teachers lack the training in phonetics and phonology to provide helpful feedback to students and correct their articulation. Secondly, high student-to-teacher ratios may prevent teachers from giving adequate attention and feedback to individual students, and limit the amount of time each student can practice speaking. Furthermore, anxiety about speaking the L2 in front of their peers may make students less willing to practice speaking, and less able to absorb corrective feedback. CAPT stands to help make pronunciation training more accessible by overcoming some of these difficulties,and to deliver the type of individualized instruction, guided by sound science and pedagogy, which many learners may not otherwise have access to.

Although much work still needs to be done to improve our understanding of how best to teach pronunciation, existing research reveals a few general considerations that must be kept in mind. First of all, it is important to note that intelligibility, and not lack of a "foreign accent", is generally considered to be the most important goal of pronunciation training (Neri et al., 2002; Derwing and Munro, 2005; Witt, 2012). Research on the impact of various types of pronunciation errors on intelligibility tends to indicate that errors on the prosodic (suprasegmental) level hinder intelligibility more than segmental errors (Anderson-Hsieh et al., 1992; Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007; Dlaska and Krekeler, 2013). For reducing these and other types of errors, perception training has been found to be very important (Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007). The importance of individualized corrective feedback is also generally acknowledged (Mehlhorn, 2005; Dlaska and Krekeler, 2013), though there is much to be learned about exactly when and how feedback can be most effective. This is the motivation behind the feedback generation module of the proposed tool (see sec. 4), which will hopefully facilitate further research on feedback in CAPT.

2.1.2 Prosody in existing CAPT systems

The viability of CAPT has been demonstrated by a variety of systems and tools that have been developed in both academic and commercial contexts. Some tools focus on overall assessment of pronunciation or fluency, while others focus on the detection and correction of individual pronunciation errors (Eskenazi, 2009); the tool developed in this work will fall into the latter category. In error-focused systems, a distinction has typically been drawn between phonemic errors, e.g. the substitution, insertion, or deletion of a segmental speech sound, and prosodic errors, such as those related to stress/accent, intonation, or rhythm (Witt, 2012). As we saw in the previous section, prosodic errors have a larger impact on intelligibility, and will be the focus of this work (see sec. 2.3 below). With this in mind, a few CAPT systems relevant to this thesis are discussed below; overviews and comparisons of these and many other systems are given by Neri et al. (2002), Eskenazi (2009), Delmonte (2011), and Witt (2012).

Both the diagnosis and feedback modules of the CAPT tool developed in this work will build to a great extent on previous work by researchers in the speech group at LORIA in Nancy, many of whom are also involved in the IFCASL project (see sec. 1.1). Their work has, on the one hand, investigated the task of automatically recognizing and segmenting learners' speech, and determining how this possibly incorrect automatic segmentation can be effectively utilized in the context of pronunciation tutoring, particularly at the prosodic level (Mesbahi et al., 2011; Orosanu et al., 2012); see sec. 3 for a discussion of how this thesis will build upon that work. Additionally, the group has developed the *Snoori* suite of software, including the PC-based WinSnoori and its partial Java port, Jsnoori (Project-Team PAROLE, 2013). These programs take as input a learner utterance, a native reference utterance, and segmentations of each, perform an acoustic comparison of the two utterances, and deliver feedback on the learner's speech in the form of e.g. annotated displays of the speech signal and spectrogram of each. Moreover, auditory feedback can be delivered thanks to the capability of resynthesizing the learner's utterance to match the pitch contour and timing of the reference, without modifying the voice quality of the utterance, such that the learner can hear the "correct" pronunciation in their own voice. The utility of such software, and especially this resynthesized feedback, for pronunciation teaching has been explored by Bonneau and Colotte (2011), who used it to assess and deliver feedback on lexical stress in L1 French speakers' pronunciation of English words. As described later in this paper, the proposed thesis will, on the one hand, build on the error detection and diagnosis functionality of Jsnoori (see sec. 3), as well as leverage its feedback generation capabilities to deliver a more diverse, and potentially more effective, range of feedback types (see sec. 4).

This work will also draw from research conducted at Carnegie Mellon University in Pittsburgh, particularly in the context of the Fluency pronunciation training system (Eskenazi and Hansma, 1998; Eskenazi et al., 2000; Probst et al., 2002) and the LISTEN project and its Reading Tutor (Duong et al., 2011; Mostow, 2012; Mostow and Aist, 1999; Sitaram et al., 2011).

In the Fluency CAPT system, particular emphasis is placed on user-adaptivity, corrective articulatory feedback, and the integration of perceptual training (e.g. listening exercises) (Eskenazi et al., 2000). As with the work at LORIA described above, the Fluency system evaluates learners' speech via comparison with that of a native reference speaker, and Probst et al. (2002) found that selecting a "golden speaker" whose voice closely matched the learner's improved learning gains. Fluency also implements an error-catching step to reject utterances which do not match the expected text (Eskenazi et al., 2000), in the same vein as that of Mesbahi et al. (2011) and Orosanu et al. (2012). Eskenazi et al. (2007) report that Fluency's commercial spin-off, NativeAccent$^{\text{TM}}$, has been shown to help real-world users significantly improve their pronunciation skills.

The Reading Tutor is not strictly a CAPT tool, as it is designed to help children develop their reading fluency in their native language. However, as it analyzes the prosody of children's read speech to measure reading fluency, and offers feedback on this prosody, it is nevertheless very relevant to CAPT and thus this thesis. In the Reading Tutor, the child's read speech is automatically segmented and compared either to a reference utterance by an adult reader, analogous to the native speaker reference in many CAPT systems, or to a generalized model of adult prosody (Duong et al., 2011). Analysis of the pitch and intensity contours of the utterances, as well as the duration of words/syllables and the spaces between them, results in an assessment of the child's overall fluency as well as identification of words which have been pronounced (in)correctly, and feedback is delivered visually in real time by revealing the text of each word as it is spoken, with properties such as the position, color, and font size of each word reflecting various aspects of the reader's prosody (Sitaram et al., 2011). This thesis will incorporate ideas and techniques from the Reading Tutor, in both its diagnosis (see sec. 3) and feedback (see sec. 4) modules.

The vast majority of CAPT systems which analyze learners' speech at the prosodic level have been developed with English as the target L2. Notable exceptions include VILLE for Swedish (Wik et al., 2009), which features both perception and production exercises for lexical stress and other prosodic errors, and Jilka and Möhler (1998)'s use of F0 contour manipulation in studying L1 English speakers' production of German. **[TODO anything else?]**

2.2 Lexical stress

When there is a typological difference between some segmental or prosodic feature(s) of a language learner's L1 compared to the target L2, there is a particular need for pronunciation training to bridge this gap. In the case of the French-German language pair, the prosodic realization of lexical stress is one feature which marks a striking difference between the languages.

Lexical stress is the phenomenon of how syllables are accentuated within a word (Cutler, 2005). This relates not to the segmental characteristics of a syllable, i.e. the speech sounds it contains, but rather to its (relative) suprasegmental properties, namely:

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

As Cutler (2005) points out, different languages make use of this suprasegmental information in different ways. In what are termed free- or variable-stress languages, such as German, Spanish, and English, it is not always possible to predict which syllable in a word will carry the stress, and therefore knowing a word requires, in part, knowing its stress pattern. This allows stress to serve a contrastive function in these languages, such that two words may share exactly the same sequence of phones and nevertheless be distinguished exclusively by their stress pattern, as is the case with *UMfahren* (to drive around) and *umFAHRen* (to run over with a car) in German. Because stress carries meaning thus, native speakers of such languages are sensitive to stress patterns, and readily able to perceive differences in stress.

However, in the so-called fixed-stress languages, stress is completely predictable, as it always falls on a certain position in the word; in Czech and Hungarian, for example, stress always falls on the word-initial syllable. Therefore, lexical stress may not be as crucial to the knowledge of a word in these languages as in the free-stress languages. Furthermore, although lexical stress is realized in these languages, the distinction between stressed and unstressed syllables may be weaker than in free-stress languages. French has often been placed into this category of fixed-stress languages, although it may be more properly considered a language without lexical stress, insofar as there is no systematic way in which speakers distinguish a certain syllable from others in the word, aside from the fact that French exhibits phrasal accent, i.e. lengthening of the final syllable in each prosodic group or phrase (Dupoux et al., 2008).

Therefore, native speakers of French may lack the sensitivity to stress patterns possessed by native speakers of German. Indeed, this has been borne out by research by Dupoux et al. (2008), who found that native French speakers are "deaf" to differences in stress patterns, such that they have great difficulty discriminating between Spanish words which contrast only at the level of stress. This difficulty should also exist for French speakers when they are presented with German words in which the stress pattern is crucial to the word's meaning, as in the minimal pair above.

2.3 Targeting lexical stress errors in CAPT

Learners of a foreign language typically make a wide variety of pronunciation errors, at both the segmental level (e.g. errors in producing certain individual phones of the target language) and the prosodic level (e.g. errors in the speaker's intonation contour or the duration of certain syllables or words). As it is not possible to address all of these in an automated system, one of the first aims of this work is to identify a single type of error which is well suited to being addressed via a CAPT system targeting French L1 learners of German as the L2.

To guide this selection, we may consider a set of three criteria that such an error must meet; similar criteria are proposed by Cucchiarini et al. (2009).

1. The error must be *produced relatively frequently* by French L1 speakers in their production of L2 German, as it would be a misuse of resources to design a system which addresses an error that is seldom made by learners (Neri et al., 2002).

2. The error must have a significant *impact on the perceived intelligibility* of the learner's speech,as the ultimate goal of the system is to help learners communicate more effectively in the L2.

3. In order for the CAPT system to provide any meaningful diagnosis and feedback, the error must lend itself to reasonably accurate and reliable *detection through automatic processing.*

Lexical stress errors fulfill all three of these criteria, and this error type has therefore been chosen as the target of the proposed CAPT tool; the remainder of this section justifies that choice.

First, as mentioned in sec. 2.1.1 above, errors related to prosody have generally been found to have a larger impact on intelligibility than segmental errors, and several studies have found lexical stress to be particularly important for comprehension in free-stress languages like English, Dutch, and our target language, German (Hirschfeld, 1994; Cutler, 2005). Secondly, we saw in sec. 2.2 that perceiving contrasts in lexical stress is notoriously difficult for native French speakers (Cutler, 2005; Dupoux et al., 2008), and given the strong link between perception and production mentioned earlier (sec. 2.1.1), this is a good indication that L1 French speakers will regularly make lexical stress errors in an L2 with free, contrastive stress, such as German. Bonneau and Colotte (2011) report that in a pilot study of L1 French speakers pronouncing English words, lexical stress was frequently misplaced by beginners; given the similarities of the lexical stress systems of English and German compared to that of French, this is another sign that we can expect such errors to be produced frequently. Finally, although much research still needs to be done on automatic detection and diagnosis of lexical stress errors (one of the main motivations behind this work; see sec. 3), recent work on this problem has shown encouraging results. As mentioned above, several existing CAPT tools incorporate treatment of lexical stress errors (e.g. Wik et al., 2009; Bonneau and Colotte, 2011, and Shahin et al. (2012) and Kim and Beutnagel (2011) have reported success in applying machine learning methods to the classification of lexical stress patterns in English words.

As lexical stress errors thus fulfill the aforementioned criteria, they will be the focus of the proposed CAPT system. The following sections describe how this thesis project will explore automatic diagnosis (sec. 3) and feedback generation (sec. 4) for this type of error.

3 Diagnosis of lexical stress errors

In order to provide learners with useful feedback on their lexical stress errors in the L2, the CAPT system must first be able to detect and diagnose such errors in a learner's utterance. This requires at least:

(a) Reasonably accurate word-, syllable- and phone-level segmentation of the learner's L2 utterance;

(b) A means of analyzing how lexical stress is realized in the prosody of the segmented utterance;

(c) A representation of how native speakers of the target language (would) realize lexical stress in the given sentence; and

(d) A way of comparing the learner's prosody to this representation.

In this section, we will examine how (a) will be achieved using forced-alignment segmentation of the learner's read-speech utterance with the corresponding text, and how problems in accuracy of the resulting segmentation can be overcome (sec. 3.1); how the lexical stress analysis in (b) can be performed by measuring the fundamental frequency (F0), duration, and energy of the relevant parts of the speech signal (sec. 3.2); and finally a variety of approaches to (c) and (d) (sec. 3.3).

3.1 Automatic segmentation of nonnative speech

Automatic segmentation, or labeling, of a recorded utterance is the task of annotating the speech signal with boundaries that demarcate individual phones, syllables, words, sentences, and/or other units of speech. Such segmentation enables analysis of a learner's L2 pronunciation, since it allows us to compare units of their speech to equivalent units in native speech.

The native and non-native read speech recordings comprising the IFCASL corpus Fauth et al., 2014; Trouvain et al., 2013 have been automatically segmented via forced alignment (see e.g. Fohr et al., 1996; Mesbahi et al., 2011). This technique requires knowledge of the text of the given utterance (known beforehand in the IFCASL case), acoustic models of the target language, and a pronunciation lexicon that describes the sequence of phones expected for each word. To account for non-native pronunciations, the lexicon has been supplemented with a lexicon of non-native variants that might be encountered for each word.

The IFCASL recordings have already been segmented at the phone and word levels, and a subset of these automatic segmentations has been manually verified. However, segmentation at the syllable level still needs to be performed. This may be accomplished based on the word- and phone-level annotations by automatically or manually determining the sounds between which syllable boundaries are expected in each sentence from the text and phonetic lexicon, automatically extracting the locations of these boundaries from the phone-level segmentation, and automatically combining those boundaries with the word-level boundaries to create a new annotation level.

The accuracy of the forced-alignment segmentation can be assessed by computing inter-annotator agreement between the automatically produced segmentation and one or more segmentations manually verified by human annotators. The team at LORIA in Nancy has already completed this evaluation for the French IFCASL sub-corpus, using the CoALT tool developed there (Fohr and Mella, 2012). In cooperation with that team, the German sub-corpus (or a subset thereof) will be evaluated in the same way. A similar evaluation will be carried out the syllable-level segmentations, a subset of which will be manually verified.

Error analysis will be performed for each boundary type, to enable identification of the types of boundaries at which the system tends (not) to make many errors. This detailed analysis will contribute to error management in the system, as described below.

Forced alignment is not a perfect method; because of the constraints put on the recognition system, the aligner will always find a match between the given text and audio, even if they do not correspond. Incorrect segmentation can lead to mistakes in error diagnosis, so CAPT systems must have a means of reducing, or at least monitoring, the amount of error introduced by inaccurate segmentation (Eskenazi, 2009). In the proposed CAPT tool, this function may be served by the development of a simple sentence- and/or word-level confidence measure. While it is very difficult to compute such a measure directly from the decoding scores of the forced aligner, it may be possible to determine which types of boundaries (e.g. between a sonorant and a vowel) the aligner typically has trouble detecting accurately, and to determine, for a given utterance, the proportion of error-prone boundaries. While a very simplistic measure, this could nevertheless provide some indication of when (not) to trust the automatic alignment, thus impacting decisions on how and whether to attempt error diagnosis (or feedback). Other error-management strategies may also be explored, such as the type of error-filtering methods described by Mesbahi et al. (2011), Bonneau et al. (2012), and Orosanu et al. (2012), in which utterances which do not correspond to the expected text are detected and rejected before alignment is attempted.

3.2 Prosodic analysis

This section will describe the features by which the system analyzes the lexical stress prosody of an utterance, be it the utterance of a learner or of a native speaker. These features relate to the three properties described in sec. 2.2, namely fundamental frequency or F0 (pitch), duration (timing), and intensity (loudness). The features computed for each property are described in the corresponding sections below.

Where possible, the diagnosis module of the CAPT tool will provide researchers control over the features used; for example, there may be an option to include all F0 and duration features but ignore intensity features.

One potential complication of this analysis that should be pointed out relates to the fact that we are here dealing exclusively with read, and not sponaneous, speech. As Cutler (2005, p. 275) remarks, "acoustic differences between stressed and unstressed syllables are relatively large in spontaneous speech. With laboratory-read materials, however, such differences do not always arise". Therefore, the task of recognizing prosodic deviations in learners' read speech may be somewhat different than the corresponding task for spontaneous speech, and this difference should be kept in mind.

As described in sec. 2.2, the fundamental frequency (F0) of an utterance, which corresponds at the perceptual level to its pitch, provides a strong signal of how lexical stress is realized in that utterance, and F0 features should therefore figure prominently in the system's prosodic analysis.

Much of the work on assessing non-native lexical stress has been conducted with English as the L2, and thus often makes the assumption that a stressed syllable should have a higher F0 than unstressed syllables (Bonneau and Colotte, 2011). In German, the F0 of a stressed syllable also tends to differ from the surrounding contour, but the difference may be positive (the stressed syllable has a higher pitch) or negative (lower pitch) (Cutler, 2005, p. 267). Therefore, features used to represent F0 may include the absolute value of the difference in average F0 between each pair of adjacent syllables in the word, or perhaps between the syllable which should carry (primary) stress and the rest of the word. To guard against unvoiced segments interfering with the F0 analysis, syllables may be represented by the vowels that form their nuclei. Relative differences may be more helpful than absolute differences. The F0 variation (range) over the entire word might be informative of whether or not the speaker failed to stress any syllable, although it would not tell us which syllable were stressed. Other features may be drawn from related work on lexical stress in learner speech, such as Bonneau and Colotte (2011).

Analysis of duration (timing) is also important for detecting stress patterns, and in this work, following Bonneau and Colotte (2011), this analysis will most likely take into account the relative durations of each syllable of the word in question, and/or of the vowel at the nucleus of each syllable. Other features may also be explored.

Research on lexical stress prosody has generally indicated that intensity is the least important of the three features, i.e. corresponds least closely to lexical stress patterns (Cutler, 2005). Indeed, existing lexical stress assessment tools may not take intensity into account, as is the case in the system described by Bonneau and Colotte (2011). However, intensity can nonetheless have an impact on the perception of lexical stress, especially in combination with pitch or duration, or both (Cutler, 2005); Therefore, the diagnosis system should ideally take intensity into account when performing its prosodic analysis. This could be as simple as computing the total energy of the part of the signal corresponding to each syllable of the word in question, although more complex measures may be explored if time allows.

3.3 Comparison of native and nonnative speech

This thesis will explore a variety of approaches to modeling the lexical stress prosody of native speech in such a way that the learner's utterance can be automatically compared to that native model. This investigation, and the creation of a CAPT tool that allows researchers to easily switch between approaches to study their effects, will be one of the primary contributions of the thesis.

The most common approach to assessing L2 prosody involves comparing a learner's utterance to a single utterance of the same sentence produced by a native speaker of the target language; this approach is taken by Bonneau and Colotte (2011) and many others (Eskenazi, 2009; Delmonte, 2011). Inspired and informed by the investigations of Probst et al. (2002), this work will examine different ways of selecting the reference speaker against which a learner's utterance will be judged, given a pool of potential references.

The most basic way of selecting a reference speaker is to choose one manually. The CAPT tool will therefore enable the learner and/or the instructor/experimenter to choose a reference from a set of available speakers, with that set potentially being constrained by one or more properties of the speaker (e.g. gender).

Another means of selecting a reference speaker would be to automatically choose a speaker whose voice resembles (or dissembles) that of the learner. By analyzing speaker-dependent features of the speech of both the learner and the reference candidates, it should be possible for the system to rank reference candidates by proximity to the learner's voice, and this should also be an option in the CAPT tool. Relevant features may include F0 mean, F0 range, duration-based features (speech tempo), spectral analysis, and/or other features informed by research on speaker identification (Shriberg et al., 2005; Reynolds and Rose, 1995, etc.). **[TODO examples of speaker ID features]**

However, when using a single native-speaker utterance for reference, even if the chosen speaker has been chosen carefully, we may be "over-fitting" to speaker- or utterance-dependent characteristics of the reference utterance that do not accurately represent the "nativeness" of the reference speech. It would therefore be advantageous not to limit the diagnosis to comparison with a single reference speaker, but to instead compare the learner's speech with a variety of native utterances. This could be accomplished by conducting a series of one-on-one comparisons, pairing the learner utterance with a different reference utterance for each comparison, and then combining the results from all the comparisons. Factors to explore in this approach might include whether the set of reference speakers should be more or less constrained (e.g. by gender), and which metrics can be used to synthesize the one-on-one comparisons into a single diagnosis.

Alternatively, the learner's utterance could perhaps be compared directly with some unified representation of all the reference utterances; for example, if we represent each reference utterance as a point in n-dimensional space, with each dimension representing a relevant feature, the references will form a cluster which can serve as a representation of the variation permissible in native speech. By plotting the learner's utterance in the same space, it could be possible to distinguish how well (or poorly) this utterance fits into that cluster, and thereby produce a diagnosis.

Finally, a different approach may be to abstract away from the reference speaker(s). In their work on assessing children's reading fluency, Duong et al. (2011) found that evaluating a child's utterance in terms of a generalized prosody model, which predicts how a given text should be uttered, yielded more accurate fluency predictions than comparing it to a reference utterance of the text in question. It would be interesting to investigate whether the same principle applies in our CAPT scenario, so if time permits, this work will explore the possibility of constructing a more general model of native lexical stress realization, and comparing the learner's utterance directly to this model instead of to one or more reference utterances. This would differ from the multiple-reference approach described above, in that while that approach limits tutoring exercises to sentences for which we have reference utterances, the general-model approach would theoretically enable the creation of exercises with arbitrary text, including sentences for which no reference utterance has been recorded. This is also generally the approach that Shahin et al. (2012) and Kim and Beutnagel (2011) followed, so machine learning methods similar to theirs may be used.

3.4 Evaluation

[TODO something about manual annotation of lexical stress errors]

4 Feedback on lexical stress errors

(Hattie and Timperley, 2007)?

Feedback is important (Neri et al., 2002)

Since our focus is on pronunciation training and not just pronunciation assessment.

Explicit FB is necessary; learners have trouble identifying their errors when simply asked to listen to what they said (Dlaska and Krekeler, 2013)

4.1 Visual feedback

(Sitaram et al., 2011)

As Neri et al. (2002) point out, waveforms and spectrograms are signal representations designed for speech researchers, not language learners, and the latter may have difficulty understanding these visualizations without the proper training.

Visualizations of the required articulators, such as those displayed in the Fluency pronunciation trainer (Eskenazi et al., 2000), may be helpful for correcting certain segmental errors, but are not likely of much use for correcting lexical stress.

4.2 Auditory feedback

(Bonneau and Colotte, 2011)

(Jilka and Möhler, 1998)?

4.3 Alternative feedback types

5 Conclusion

This proposal has outlined a thesis project in Computer-Assisted Pronunciation Training for native French speakers learning German as a foreign language. The project will focus on lexical stress errors, namely methods for their automatic diagnosis (sec. 3) and for the automatic generation of feedback on such errors (sec. 4). These methods will be combined into a prototype CAPT tool, apparently the first of its kind for this L1/L2 language pair, which will facilitate research on the efficacy of different feedback types for such errors, and could eventually be incorporated into a full CAPT system for German.

References

Anderson-Hsieh, Janet, Ruth Johnson, and Kenneth Koehler (1992). “The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure”. In: *Language Learning* 42.4, pp. 529–555 (cit. on p. 3).

Bonneau, Anne and Vincent Colotte (2011). “Automatic Feedback for L2 Prosody Learning”. In: *Speech and Language Technologies.* Ed. by Ivo Ipsic. 1977. InTech (cit. on pp. 4, 7, 9, 10, 12).

Bonneau, Anne, Dominique Fohr, Irina Illina, Denis Jouvet, Odile Mella, Larbi Mesbahi, and Luiza Orosanu (2012). “Gestion d’erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d’une langue seconde”. In: *Traitement Automatique des Langues* 53, pp. 129–154 (cit. on p. 9).

Cucchiarini, Catia, Ambra Neri, and Helmer Strik (2009). “Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback”. In: *Speech Communication* 51.10, pp. 853–863 (cit. on p. 6).

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. 5, 7, 9, 10).

Delmonte, Rodolfo (2011). “Exploring Speech Technologies for Language Learning”. In: *Speech and Language Technologies.* Ed. by Ivo Ipsic. InTech (cit. on pp. 2, 3, 10).

Derwing, Tracey M and Murray J Munro (2005). “Second Language Accent and Pronunciation Teaching: A Research-Based Approach”. In: *TESOL Quarterly* 39.3, pp. 379–397 (cit. on pp. 1, 3).

Dlaska, Andrea and Christian Krekeler (2013). “The short-term effects of individual corrective feedback on L2 pronunciation”. In: *System* 41.1, pp. 25–37 (cit. on pp. 3, 12).

Duong, Minh, Jack Mostow, and Sunayana Sitaram (2011). “Two methods for assessing oral reading prosody”. In: *ACM Transactions on Speech and Language Processing* 7.212, pp. 1–22 (cit. on pp. 4, 5, 11).

Dupoux, Emmanuel, Núria Sebastián-Gallés, Eduardo Navarette, and Sharon Peperkamp (2008). “Persistent stress ’deafness’: The case of French learners of Spanish”. In: *Cognition* 106, pp. 682–706 (cit. on pp. 6, 7).

Eskenazi, Maxine (2009). “An overview of spoken language technology for education”. In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 2, 3, 8, 10).

Eskenazi, Maxine and Scott Hansma (1998). “The Fluency pronunciation trainer”. In: *Proc. of Speech Technology in Language Learning*, pp. 77–80 (cit. on p. 4).

Eskenazi, Maxine, Yan Ke, Jordi Albornoz, and Katharina Probst (2000). “The Fluency Pronunciation Trainer: Update and user issues”. In: *Proc. of InSTIL 2000, Dundee* (cit. on pp. 4, 12).

Eskenazi, Maxine, Angela Kennedy, Carlton Ketchum, Robert Olszewski, Garrett Pelton, Forbes Ave, and Pittsburgh Pa (2007). “The NativeAccent(TM) pronunciation tutor: measuring success in the real world”. In: *SLaTE*, pp. 124–127 (cit. on p. 4).

Fauth, Camille, Anne Bonneau, and Frank Zimmerer (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. 1, 8).

Fohr, Dominique and Odile Mella (2012). "CoALT: A Software for Comparing Automatic Labelling Tools." In: *LREC*, pp. 325–332 (cit. on p. 8).

Fohr, Dominique, JF Mari, and Jean Paul Haton (1996). "Utilisation de modèles de Markov pour l'étiquetage automatique et la reconnaissance de BREF80". In: *Journées d'Etude de la Parole* (cit. on p. 8).

Hattie, J and H Timperley (2007). "The power of feedback". In: *Review of Educational Research* 77.1, pp. 81–112 (cit. on p. 11).

Hirschfeld, Ulla and Jürgen Trouvain (2007). "Teaching prosody in German as foreign language". In: *Non-Native Prosody: Phonetic Description and Teaching Practice*. Ed. by Jürgen Trouvain and Ulrike Gut. Walter de Gruyter, pp. 171–187 (cit. on p. 3).

Hirschfeld, Ursula (1994). *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Institut für Phonetik, JW Goethe-Universität (cit. on p. 7).

Jilka, M and G Möhler (1998). "Intonational foreign accent: speech technology and foreign language teaching". In: *.... ESCA Workshop on Speech Technology in ...* (Cit. on pp. 5, 12).

Kim, Yeon-Jun and Mark C Beutnagel (2011). "Automatic assessment of american English lexical stress using machine learning algorithms." In: *SLaTE*, pp. 93–96 (cit. on pp. 7, 11).

Mehlhorn, G (2005). "Learner autonomy and pronunciation coaching". In: *Proceedings of the Phonetics Teaching and Learning Conference, University College London* (cit. on p. 3).

Mesbahi, Larbi, Denis Jouvet, Anne Bonneau, and Dominique Fohr (2011). "Reliability of non-native speech automatic segmentation for prosodic feedback." In: *SLaTE* (cit. on pp. 4, 8, 9).

Mostow, Jack (2012). "Why and how our automated reading tutor listens". In: *International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)* (cit. on p. 4).

Mostow, Jack and Gregory Aist (1999). "Giving help and praise in a reading tutor with imperfect listening–because automated speech recognition means never being able to say you're certain". In: *CALICO journal* (cit. on p. 4).

Neri, A., C. Cucchiarini, H. Strik, and L. Boves (2002). "The pedagogy-technology interface in computer assisted pronunciation training". In: *Computer Assisted Language Learning* (cit. on pp. 2, 3, 6, 11, 12).

Orosanu, Luiza, Denis Jouvet, Dominique Fohr, Irina Illina, and Anne Bonneau (2012). "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* (cit. on pp. 4, 9).

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 pp. 4, 10).

Project-Team PAROLE (2013). *Activity Report 2013.* Tech. rep. Nancy: LORIA (cit. on p. 4).

Reynolds, Douglas A. and Richard C. Rose (1995). "Robust text-independent speaker identification using Gaussian mixture speaker models". In: *IEEE Transactions on Speech and Audio Processing* 3.1, pp. 72–83 (cit. on p. 10).

Shahin, Mostafa Ali, Beena Ahmed, and Kirrie J. Ballard (2012). "Automatic classification of unequal lexical stress patterns using machine learning algorithms". In: *2012 IEEE Spoken Language Technology Workshop (SLT).* IEEE, pp. 388–391 (cit. on pp. 7, 11).

Shriberg, E., L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke (2005). "Modeling prosodic feature sequences for speaker recognition". In: *Speech Communication* 46.3-4, pp. 455–472 (cit. on p. 10).

Sitaram, S, J Mostow, Y Li, A Weinstein, D Yen, and J Valeri (2011). "What visual feedback should a reading tutor give children on their oral reading prosody?" In: *SLaTE* (cit. on pp. 4, 5, 12).

Trouvain, Jürgen, Yves Laprie, and Bernd Möbius (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 pp. 1, 8).

Wik, P, R Hincks, and JB Hirschberg (2009). "Responses to Ville: A virtual language teacher for Swedish". In: (cit. on pp. 5, 7).

Witt, Silke M (2012). "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)*, pp. 1–8 (cit. on pp. 1–3).