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

Towards a CAPT system for French learners of German

Anjana Sofia Vakil

A thesis submitted toward the degree of

Master of Science

in Language Science and Technology

Prepared under the supervision of

Prof. Dr. Bernd Möbius

Dr. Jürgen Trouvain

UNIVERSITÄT DES SAARLANDES

Saarland University

Department of Computational Linguistics & Phonetics

11 December, 2014

Anjana Sofia Vakil
anjanav@coli.uni-saarland.de
Automatic diagnosis and feedback for lexical stress errors in non-native speech
11 December, 2014
Supervisors: Prof. Dr. Bernd Möbius and Dr. Jürgen Trouvain

Saarland University
Department of Computational Linguistics & Phonetics
Fachrichtung 4.7 Allgemeine Linguistik
Postfach 15 11 50
66041 and Saarbrücken

Typeset using LaTeX 2_ε. Style adapted from the *Clean Thesis* template developed by Ricardo Langner (http://cleanthesis.der-ric.de/).

Declaration

I hereby declare that this thesis, presented here toward the completion of the Master of Science degree, is my own original work. All material contained herein that has been taken from other sources is acknowledged as such, without exception.

Saarbrücken, 11 December, 2014

Anjana Sofia Vakil

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Abstract (different language)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgement

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

List of Figures

List of Tables

1 Introduction

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 very differently in the French language.

Computer-Assisted Pronunciation Training (CAPT) systems have the potential to automatically provide highly individualized analysis of such learner errors, as well as feedback on how to correct them, and thus to help learners achieve more intelligible pronunciation in the target language (Witt, 2012). The thesis project proposed here 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 Saarland University (Saarbrücken, Germany) and LORIA (Nancy, France).

The goal of the IFCASL project is to take initial steps toward the development of a CAPT system targeting native (L1) French speakers learning German as a foreign language (L2), as well as 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. Furthermore, 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

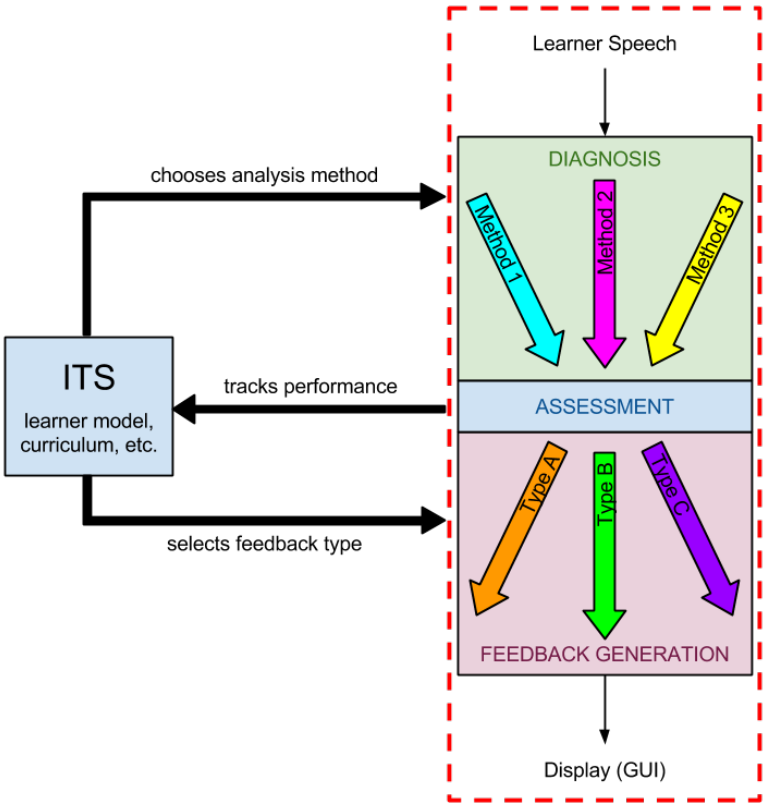

Figure 1.1: Conceptual diagram of the prototype lexical stress CAPT tool (demarcated by dotted line) and its possible function in the context of a more comprehensive Intelligent Tutoring System.

intended outcome of these investigations is a prototype CAPT tool, illustrated in fig. 1.1, which can diagnose lexical stress errors in different ways and present learners with different types of feedback on these errors.

Learners should be able to interact with the tool and interpret its feedback independently, i.e. without the assistance of a human instructor at their side. Researchers should be able to 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 Thesis overview

Chapter 2 places this thesis in the context of existing research on CAPT, and motivates its specific focus on lexical stress errors. Chapter 4 outlines the techniques to be explored for diagnosing lexical stress errors in learners' speech via automatic processing of acoustic correlates of these errors in a spoken utterance. Chapter 5 describes the multimodal feedback types the system will aim to deliver, based on the analysis described in the previous section. **??** summarizes the proposal and the aims of the thesis.

Background and related work

2

2.1 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. 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 foreign language education, 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.

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 (Neri et al., 2002; 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 chapter 5), which is intended to facilitate research on CAPT feedback.

2.2 Computer-Assisted Pronunciation Training

Much of the recent interest in CAPT stems from its potential to deliver the type of individualized pronunciation instruction, guided by sound science and pedagogy, which many learners may not have access to in the classroom, as just discussed. 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 focus on overall assessment of pronunciation or fluency, and others 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, word-prosodic errors have a larger impact on intelligibility, and will be the focus of this work (see sec. 2.4 below). With this in mind, a few prosody-aware 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 chapter 4 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 chapter 4), as well as leverage its feedback generation capabilities to deliver a more diverse, and potentially more effective, range of feedback types (see chapter 5).

This work will also draw from research on two systems developed at Carnegie Mellon University. The first of these, the Fluency pronunciation trainer (Eskenazi and Hansma, 1998; Eskenazi et al., 2000), is a CAPT system placing particular emphasis on user-adaptivity, corrective articulatory feedback, and the integration of perceptual training (e.g. listening exercises). 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™, has been shown to help real-world users significantly improve their pronunciation skills.

A second CMU system, the Project LISTEN Reading Tutor (Mostow, 2012) is designed to help children develop reading fluency in their native language. To that end, it analyzes the prosody of children's read speech to measure reading fluency, and offers feedback on this prosody. 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) report better performance using the generalized model. Analysis of the pitch and intensity contours of the utterance(s), as well as the duration of words/syllables and the pauses 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). Ideas and techniques from the Reading Tutor will influence both the diagnosis (see chapter 4) and feedback (see chapter 5) modules of the proposed CAPT tool.

The vast majority of CAPT systems which analyze learners' speech at the prosodic level have been developed with English as the target L2, and relatively little work has been done on German. In a notable exception particularly relevant to this thesis, Bissiri et al. (2006; 2009) found that L1 Italian speakers' realizations of lexical stress in German improved when they were allowed to listen to prosodically-modified recordings of their own speech and that of native speakers (see sec. 5.2). Jilka and Möhler's (1998) use of F0 contour manipulation in studying L1 English speakers' production of German represents another exploration of speech technology applications for German instruction. Language-independent tools have also been developed, such as WinPitch LTL (Martin, 2004), which enables speech signal visualization of prosodic features such as pitch contours as well as manipulation of prosody and comparison to reference utterances, with the intent that a human instructor will guide the learner in using the software and interpreting the visualizations.

2.3 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 an uttered 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 lexical 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. Furthermore, in German, misplaced stress has been shown to disrupt understanding of a word or utterance even in cases where there is no stress-based minimal pair (Hirschfeld, 1994), supporting the theory that speakers of free-stress languages rely to a large extent on stress information in the recognition of spoken words (Cutler, 2005).

However, in the so-called fixed-stress languages, stress is completely predictable, as it always falls on a certain position in the word; in French, for example, stress is fixed on the word-final syllable, while in Czech and Hungarian, stress always falls on the initial syllable. 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.

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.

2.4 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 feasible to address all of these in a prototype CAPT tool, one of the first aims of this work is to identify a single type of error which is well suited to being addressed via CAPT for L1 French/L2 German.

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). First, 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 addressing an error seldom made by learners (Neri et al., 2002). Second, 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. Third, 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*. As illustrated in fig. 2.1, the best error to target with the CAPT system will fulfill all of these criteria, rather than only one or two of the three.

Figure 2.1: Criteria for selecting errors to target in a CAPT system.

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.

2.4.1 Impact on intelligibility

First, as mentioned in sec. 2.1 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).

2.4.2 Frequency of production

Secondly, we saw in sec. 2.3 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, 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.

2.4.3 Feasibility of automatic detection

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 chapter 4), 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 (chapter 4) and feedback generation (chapter 5) for this type of error.

2.5 Summary

3 System overview?

4 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 given 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, and how problems in accuracy of the resulting segmentation can be overcome (sec. 4.1); 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. 4.2); and finally a variety of approaches to (c) and (d) (sec. 4.3).

4.1 Automatic segmentation of nonnative speech

4.1.1 Segmentation via forced alignment

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 (Fohr et al., 1996; Mesbahi et al., 2011). This technique requires the expected text of the given utterance, 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 is 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.

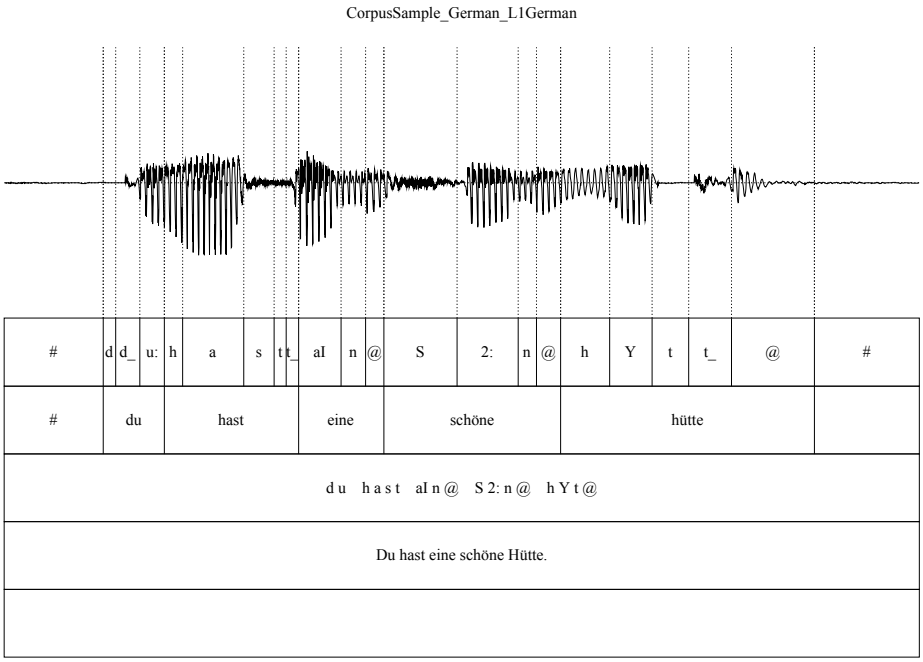

Figure 4.1: An example of a German utterance that has been segmented at the phone level (first row) and word level (second row). The third row contains the canonical (expected) native pronunciation of each word in the sentence, while the fourth row contains the written sentence of which the utterance is a reading.

4.1.2 Evaluation of segmentation accuracy

The accuracy of the forced-alignment segmentation can be assessed by computing inter-annotator agreement between the automatically produced segmentation and one or more manually-verified segmentations. The team at LORIA in Nancy has already completed this evaluation for the French IFCASL sub-corpus using the CoALT tool (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 for the syllable-level segmentations, a subset of which will be manually verified.

4.1.3 Coping with segmentation errors

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 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 from the aforementioned accuracy evaluation which types of boundaries (e.g. between a sonorant and a vowel) the aligner typically has trouble detecting accurately, and then to calculate, 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.

4.2 Analysis of word prosody

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. 2.3, namely duration (timing), fundamental frequency or F0 (pitch), 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.

4.2.1 Duration

Analysis of duration (timing) is extremely important for detecting stress patterns; indeed, syllable duration may be the most important acoustic correlate for lexical stress in German (Dogil and Williams, 1999). In this work, duration analysis will figure prominently, and following Bonneau and Colotte (2011) will most likely take into account the relative duration of each syllable of the word in question, and/or of the vowel at the nucleus of each syllable.

4.2.2 Fundamental frequency

As described in sec. 2.3, the fundamental frequency (F0) of an utterance, which corresponds at the perceptual level to its pitch, also provides a strong signal of how lexical stress is realized in that utterance, and F0 features should therefore also contribute to 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 between syllables may be more helpful than absolute differences. The F0 variation (range) over the entire word might also be informative of whether or not the speaker failed to stress any syllable.

4.2.3 Intensity

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.

4.3 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.

4.3.1 Using a single reference speaker

The most common approach to assessing L2 prosody involves comparing a learner's utterance to the same utterance produced by a native speaker of the target language; this approach is taken by Bonneau and Colotte (2011) and others.

Manually selecting a reference

The most basic way of selecting a reference speaker is to choose one manually. As a type of baseline, 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).

Automatically selecting a reference

Another means of selecting a reference speaker would be to automatically choose a speaker whose voice resembles that of the learner (Probst et al., 2002). By analyzing speaker-dependent features of the speech of each reference candidate and of the learner – possibly in their L1 (French) as well as the L2 (German) – it should be possible for the system to rank reference candidates by proximity to the learner's voice. Relevant features may include F0 mean/range as well as spectral and duration-based features.

4.3.2 Using multiple reference speakers

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.

4.3.3 Using no reference speaker

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 theoretically enable the creation of exercises with arbitrary text, including sentences for which no reference utterance has been recorded. Possibilities for generalized lexical stress modeling include using word-prosody predictions from a text-to-speech synthesizer such as MARY (Schröder and Trouvain, 2003), as well as classification-based machine learning approaches such as those used by Shahin et al. (2012) and Kim and Beutnagel (2011) to categorize English words based on their stress patterns.

As this last diagnostic approach, using generalized lexical stress modeling, is the one which has been least explored in CAPT research, it will be the first priority for this thesis work after the baseline approach (manually selecting a single reference speaker) has been implemented. The next highest priority will be comparing the learner's speech to multiple reference speakers, followed by automatically selecting a reference speaker to match the learner's voice; these approaches will only be explored as time allows.

4.4 Evaluation

Lexical stress errors in the manually-annotated subset of the IFCASL corpus have not been explicitly labeled. We can assume that the utterances from L1 German speakers exhibit only correct German stress patterns, but a subset of the L1 French utterances will need to be annotated for lexical stress errors. This labeled data will be needed to assess the accuracy

of the various error diagnosis methods which will be explored, and potentially to train classifiers to recognize correctly and incorrectly stressed words.

4.5 Summary

5 Feedback on lexical stress errors

Since the focus of this thesis is on pronunciation training, not pronunciation assessment (see sec. 2.2), feedback on the errors diagnosed via the methods described in chapter 4 will be an important component of the proposed CAPT tool. As mentioned in sec. 2.1, the particular importance of corrective feedback in pronunciation training is generally acknowledged, though much remains to be learned about when and how feedback can be most effective. Therefore, one aim of this thesis is the creation of a feedback generation module for the lexical stress CAPT tool which will offer a variety of possible feedback types, and a Graphical User Interface (GUI) allowing a researcher or instructor to easily switch between feedback types. While it is outside the scope of the thesis to carry out in vivo studies with learners to determine which feedback types are most effective in which situations, the tool will hopefully facilitate such studies going forward.

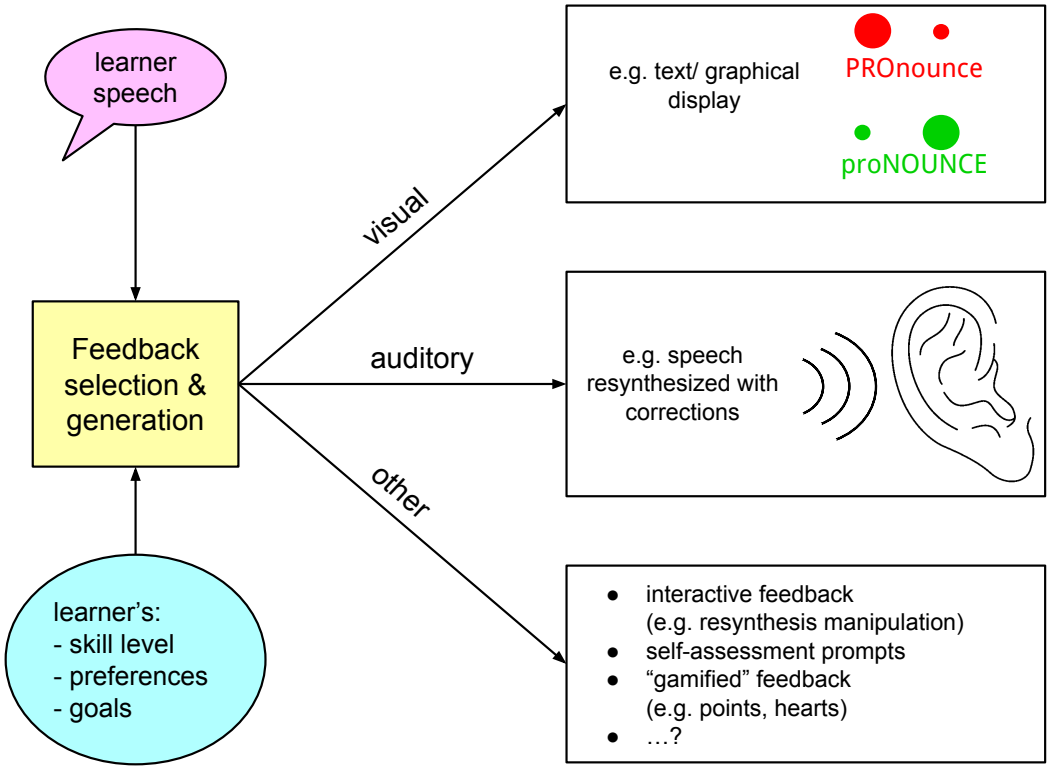

Figure 5.1: Delivery of prosody feedback in different modalities.

5.1 Visual feedback

5.1.1 Visualizations of the speech signal

In several existing CAPT tools, the learner is presented with relatively direct visualizations of the speech signal, such as its waveform (oscillogram) and spectrogram, often with overlays highlighting perceptually relevant properties such as the pitch contour and durations of various parts of the utterance. However, 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. To research whether this conjecture holds, these direct visualizations must be compared with alternatives in user studies with learners; several options for such alternative visualizations are explored in this section.

5.1.2 Graphical representations of prosody

One type of alternative would be a more abstract graphical representation of the lexical stress pattern in the native reference speaker and/or the learner's speech. Classroom materials for pronunciation instruction sometimes represent lexical stress patterns using dots or other shapes, one for each syllable, whose relative sizes indicate each syllable's prominence in the word (Hirschfeld and Reinke, 1998). This type of visualization would be relatively simple to implement, given that the reference or learner utterance can be classified into one of a set of stress patterns (Kim and Beutnagel, 2011; Shahin et al., 2012). It would also be possible to map the acoustic features of each syllable in the utterance(s) to graphical features of the representative shape, e.g. using size to represent duration, vertical position to represent F0, and darkness to represent intensity. To facilitate studies on which mappings, if any, make this feedback useful to the learner, the researcher-facing GUI should offer control over the different possible mappings.

5.1.3 Stylized text

This is essentially the approach used by Sitaram et al. (2011), though they modify the text of each word instead of a more abstract visual representation. As text stylization is also often used in pronunciation instruction materials (Behme-Gissel, 2005; Hirschfeld et al., 2007), it would be logical for the CAPT tool to offer text stylization as a feedback option. As with the shapes mentioned above, and following Sitaram et al. (2011), it would be interesting to explore the possible mappings between acoustic features and properties of the text of each syllable (e.g. size, weight, underlining/decoration, etc.), with these mappings controllable by the researcher via the GUI.

5.1.4 Other

Given some visual representation of the learner's utterance, be it textual or more abstract, visual feedback should also be given on what the learner can do to improve their lexical stress realization. Bonneau and Colotte (2011) deliver such feedback in the F0 dimension by displaying arrows which indicate whether the user should raise or lower the pitch of a given syllable to make their realization more like that of the reference speaker, and this is one option

for the CAPT tool. Another might be the use of animation to transform the visualization of the learner's (incorrectly realized) utterance into a corresponding visualization of the correct realization, e.g. by growing or shrinking the size of the dot or text for each syllable to visualize the desired change in duration, or showing it moving up or down to convey the desired change in pitch.

Implementation of at least one visual feedback type will be of high priority in this work. Stylized text and graphical representations will be explored first. If time allows, animation will be added to convey corrective feedback to the learner.

5.2 Auditory feedback

In foreign language classrooms, feedback on correct pronunciation is often given implicitly by allowing the learner to listen to a native speaker's production of the target utterance and/or a recording of their own production. However, previous work on delivering lexical stress feedback (see sec. 2.2) has revealed that learners seem to benefit more from prosodically modified implicit feedback, either in the form of a learner utterance modified to reflect the "correct" prosody of a native reference utterance (Bonneau and Colotte, 2011), or a native utterance modified to place exaggerated emphasis on the stressed syllable (Bissiri et al., 2006; Bissiri and Pfitzinger, 2009).

At least one type of audio feedback type will be implemented in the CAPT tool, with the highest-priority option being prosodic modification of the learner's utterance to match a single, manually-selected reference utterance, following Bonneau and Colotte (2011); Jsnoori (Project-Team PAROLE, 2013) will be used to perform this modification. If a generalized lexical stress model is successfully integrated into the diagnostic module (see sec. 4.3), the next highest-priority task will be performing prosodic modification of the learner's utterance based on this model. Emphasizing stressed syllables in the native reference utterance(s) will be of lowest priority.

5.3 Alternative feedback types

Other options, which will only be explored if time allows, include (in order of priority) feedback encouraging self-assessment and self-correction, metalinguistic feedback, and interactivity. Self-assessment and self-correction can be encouraged by presenting learners with targeted questionnaires before delivering diagnosis and feedback, e.g. asking learners to listen to their utterance and assess whether they have placed stress on the correct syllable, or asking how the speaker of an incorrect production could have realized stress properly ("By making the first syllable longer", etc.). Metalinguistic feedback, e.g. reminding learners of the stress rule(s) affecting the target utterance, could be delivered either visually (e.g. text displayed on the screen), auditorily (e.g. playback of an instructor's voice), or both. Interactivity could be achieved by allowing learners to interact with the resynthesis component to modify the prosody of their utterance, as is done in WinPitch LTL (Martin, 2004). By allowing researchers to easily control which of these feedback options to present

to the learner, the tool could facilitate research into the effects of alternative feedback types such as these, which have not yet been adequately studied in CAPT.

5.4 Summary

6 Conclusion and outlook

6.1 Thesis summary

6.2 Future work

Bibliography

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. 5).

Behme-Gissel, Helma (2005). *Deutsche Wortbetonung: ein Lehr- und Übungsbuch*. Iudicium (cit. on p. 20).

Bissiri, Maria Paola and Hartmut R. Pfitzinger (2009). "Italian speakers learn lexical stress of German morphologically complex words". In: *Speech Communication* (cit. on pp. 7, 20).

Bissiri, Maria Paola, Hartmut R. Pfitzinger, and Hans G. Tillmann (2006). "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis". In: *Proceedings of the 11th Australian International Conference on Speech Science & Technology* (cit. on pp. 7, 20).

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. 6, 9, 10, 15, 16, 20).

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. 15).

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. 8).

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. 7–9, 15, 16).

Delmonte, Rodolfo (2011). "Exploring Speech Technologies for Language Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on p. 6).

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 p. 5).

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 p. 5).

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. Walter de Gruyter, pp. 273–334 (cit. on p. 15).

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. 7, 17).

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. 8, 9).

Eskenazi, Maxine (2009). “An overview of spoken language technology for education”. In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 6, 14).

Eskenazi, Maxine and Scott Hansma (1998). “The Fluency pronunciation trainer”. In: *Proc. of Speech Technology in Language Learning*, pp. 77–80 (cit. on p. 6).

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 p. 6).

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. 6).

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, 13).

Fohr, Dominique and Odile Mella (2012). “CoALT: A Software for Comparing Automatic Labelling Tools.” In: *LREC*, pp. 325–332 (cit. on p. 14).

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. 13).

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. 5).

Hirschfeld, Ursula (1994). *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Institut für Phonetik, JW Goethe-Universität (cit. on pp. 8, 9).

Hirschfeld, Ursula and Kerstin Reinke (1998). *Phonetik Simsalabim: Ein Übungskurs für Deutschlernender (Begleitbuch)*. Langenscheidt (cit. on p. 19).

Hirschfeld, Ursula, Christian Keß ler, Barbara Langhoff, Kerstin Reinke, Annemargret Sarnow, Lothar Schmidt, and Eberhard Stock (2007). *Phonothek intensiv: Aussprachetraining*. Ed. by Ursula Hirschfeld, Kerstin Reinke, and Eberhard Stock. Langenscheidt (cit. on p. 20).

Jilka, M and G Möhler (1998). “Intonational foreign accent: speech technology and foreign language teaching”. In: . . . . *ESCA Workshop on Speech Technology in . . .* (Cit. on p. 7).

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. 10, 17, 19).

Martin, Philippe (2004). “WinPitch LTL II, a multimodal pronunciation software”. In: *InSTIL/ICALL Symposium 2004* (cit. on pp. 7, 21).

Mehlhorn, G (2005). “Learner autonomy and pronunciation coaching”. In: *Proceedings of the Phonetics Teaching and Learning Conference, University College London* (cit. on p. 5).

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. 6, 13, 15).

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. 7).

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. 5, 6, 8, 19).

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. 6, 15).

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. 6, 16).

Project-Team PAROLE (2013). *Activity Report 2013*. Tech. rep. Nancy: LORIA (cit. on pp. 6, 20).

Schröder, Marc and Jürgen Trouvain (2003). “The German text-to-speech synthesis system MARY: A tool for research, development and teaching”. In: *International Journal of Speech Technology* 6, pp. 365–377 (cit. on p. 17).

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. 10, 17, 19).

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. 7, 20).

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, 13).

Wik, P, R Hincks, and JB Hirschberg (2009). “Responses to Ville: A virtual language teacher for Swedish”. In: (cit. on p. 10).

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, 5, 6).