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 and native-like 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 sensitive to the lexical stress patterns of German and develop the ability to accurately realize these patterns in their speech.

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.

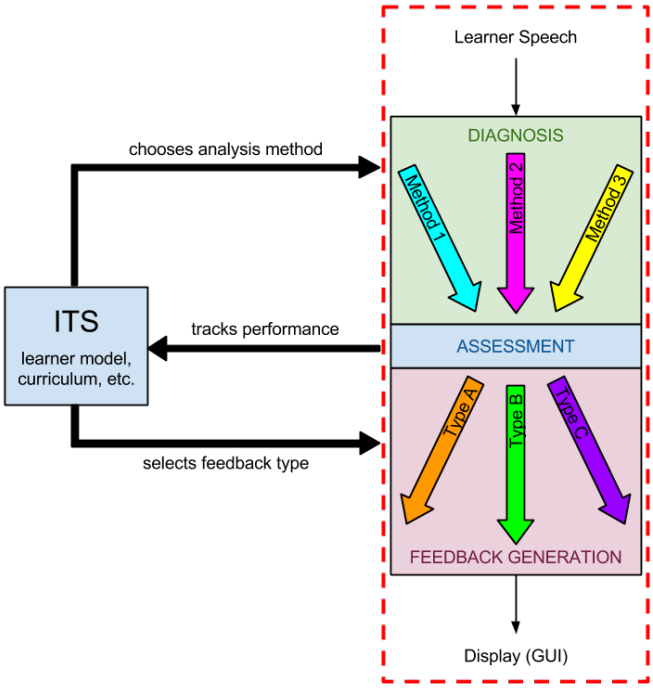

Figure 1: Conceptual diagram of the prototype lexical stress CAPT tool (demarcated by dotted line) and its possible function in the context of a more comprehensive Intelligent Tutoring System.

The intented outcome of these investigations is a prototype CAPT tool, illustrated in fig. 1, which 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

(Eskenazi, 2009; Delmonte, 2011; Witt, 2012)

2.1.1 Pronunciation in foreign language education

The difficulties posed by including pronunciation in the foreign language classroom curriculum will be discussed in this section, leading to the conclusion that CAPT can help make pronunciation training more accessible by overcoming some of these difficulties (e.g. teacher-to-student ratio). Relevant findings from a variety of works on pronunciation teaching in the classroom will be presented (Derwing and Munro, 2005; Dlaska and Krekeler, 2013; Hirschfeld and Trouvain, 2007; Mehlhorn, 2005).

2.1.2 Computer-based and intelligent tutoring systems?

This section would serve as a domain-independent overview of CBT and ITS, and the advantages such systems can bring when deployed in schools or used individually.

2.1.3 Prosody in existing CAPT systems

This section will describe a selection of related CAPT systems and tools, and how these tools have analyzed and offered feedback on speech prosody.

Both the diagnosis and feedback modules of the CAPT tool will build to a great extent on work conducted by the speech group at LORIA in Nancy (Bonneau and Colotte, 2011; Fohr et al., 1996; Fohr and Mella, 2012; Mesbahi et al., 2011; Orosanu et al., 2012). The Jsnoori/WinSnoori software (Project-Team PAROLE, 2013) which this group has developed will be instrumental in the construction of the CAPT tool.

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; 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; Weber and Bali, 2010). The latter may not strictly fall into the category of CAPT

systems, but 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 this thesis.

Other systems mentioned by Eskenazi (2009), Delmonte (2011), and Witt (2012) may also be briefly described, including tools developed at KTH (Hincks, 2002; Hincks and Edlund, 2009) to teach English prosody.

2.2 Lexical stress

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 X and Y 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. (Peperkamp and Dupoux, 2002; Dupoux et al., 2001; Dupoux et al., 2008), which demonstrated that native French speakers were "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 therefore 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. The best error to target with the CAPT system will fulfill all of these criteria, rather than only one or two of the three. First, the error must be produced with a some degree of frequency 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. Secondly, the given 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, an error which is commonly made but nevertheless does not impede understanding of the learner's L2 speech, and thus does not hinder communication in the L2, is not an ideal target. Finally, in order for the CAPT system to provide any meaningful diagnosis of and feedback on the error, it must lend itself to reasonably accurate and reliable detection through automatic processing.

This thesis proposes that lexical stress errors are a strong candidate for treatment via CAPT, and will therefore be the focus of the prototype system which will be developed. The remainder of this section justifies the targeting of lexical stress errors by describing how this type of error fulfills the aforementioned criteria.

(Warren et al., 2009)

(Magen, 1998)

Stress errors may affect perception of segmental errors; errors in stressed syllables are more noticeable (Cutler, 2005)

(Cutler, 2005)

(Peperkamp and Dupoux, 2002; Dupoux et al., 2001; Dupoux et al., 2008)

(Engwall, 2012; Delmonte, 2011)

(Bonneau and Colotte, 2011)

(Shahin et al., 2012; Kim and Beutnagel, 2011)

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) is 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 (section 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 (section 3.2); and finally a variety of approaches to (c) and (d), and the advantages and difficulties of each (section 3.3).

3.1 Automatic segmentation of nonnative speech

By way of introduction, this section will begin with a brief explanation of the task of automatically segmenting, or labeling, a non-native speaker's utterance, and some of the difficulties involved.

3.1.1 Segmentation via forced alignment

This section will briefly summarize the forced-alignment segmentation method (Fohr et al., 1996; Mesbahi et al., 2011, etc.), describe the data on which the relevant acoustic models were trained (if possible, the models will be retrained on or adapted to the IFCASL data), and point out the inclusion of a lexicon of non-native variants (which may be extracted from the IFCASL data).

3.1.2 Evaluation of system accuracy

In this section, the accuracy of the forced-alignment segmentation will be assessed by computing inter-annotator agreement between the automatically produced segmentation and one or more segmentations manually verified by human annotators. This may be conducted in cooperation with the team at LORIA in Nancy, and will make use of the CoALT tool developed there (Fohr and Mella, 2012). 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 in section 3.1.3.

3.1.3 Coping with segmentation errors

This section will explain how the system attempts to reduce the amount of error introduced by inaccurate segmentation. This may involve implementation of the error-triage methods described by Mesbahi et al. (2011), Bonneau et al. (2012), and Orosanu et al. (2012). Other strategies may also be explored, such as the development of a type of sentence- and/or word-level confidence measure based on the boundary error rates found in section 3.1.2.

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 section 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.

3.2.1 Fundamental frequency

As described in section 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. In general F0 will be computed using a standard pitch-tracking algorithm, such as that implemented in the Praat[1] software.

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,

[1] http://www.praat.org/

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).

3.2.2 Duration

Following Bonneau and Colotte (2011), timing 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 be drawn from related work.

3.2.3 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.

3.3 Comparison of native and nonnative speech

This section 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.

3.3.1 Using a single reference speaker

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 is the approach taken by Bonneau and Colotte (2011) and others.

Inspired and informed by the investigations of Probst et al. (2002), this section will examine different ways of selecting the reference speaker against which a learner's utterance will be judged, given a pool of potential references.

3.3.2 Manually selecting a reference

The most basic way of selecting a reference speaker is to manually specify which speaker should be used for comparison. 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).

3.3.3 Automatically selecting a reference

A more interesting means of selecting a reference speaker would be to automatically choose a speaker whose voice resembles to a greater or lesser degree the voice 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.).

3.3.4 Using multiple reference speakers

When using a single native-speaker utterance for reference, even if the chosen speaker's voice closely resembles that of the learner, 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.

3.3.5 Using no reference speaker?

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 in section 3.3.4 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 Summary

4 Feedback on lexical stress errors

(Hattie and Timperley, 2007)

4.1 Visual feedback

4.1.1 Stylized text

4.1.2 Graphical representations of prosody

4.1.3 Visualizations of the speech signal

4.2 Auditory feedback

4.2.1 Enhanced reference utterance

4.2.2 Resynthesized learner speech

4.3 Alternative feedback types

4.3.1 Metalinguistic feedback

4.3.2 Interactive feedback

4.3.3 Implicit feedback

4.4 Summary

5 Conclusion

References

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. 3, 5, 7, 8).

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. 7).

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. 4, 5, 7, 8).

Delmonte, Rodolfo (2011). "Exploring Speech Technologies for Language Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on pp. 3–5).

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 p. 3).

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. 3, 9).

Dupoux, Emmanuel, Sharon Peperkamp, and Núria Sebastián-Gallés (2001). "A robust method to study stress "deafness"". In: *The Journal of the Acoustical Society of America* 110.3, pp. 1606–1618 (cit. on pp. 4, 5).

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. 4, 5).

Engwall, Olov, ed. (2012). *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*. Stockholm, Sweden: KTH, Computer Science and Communication (cit. on p. 5).

Eskenazi, Maxine (2009). "An overview of spoken language technology for education". In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 3, 4).

Eskenazi, Maxine and Scott Hansma (1998). "The Fluency pronunciation trainer". In: *Proc. of Speech Technology in Language Learning*, pp. 77–80 (cit. on p. 3).

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 p. 1).

Fohr, Dominique and Odile Mella (2012). "CoALT: A Software for Comparing Automatic Labelling Tools." In: *LREC*, pp. 325–332 (cit. on pp. 3, 6).

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 pp. 3, 6).

Hattie, J and H Timperley (2007). "The power of feedback". In: *Review of Educational Research* 77.1, pp. 81–112 (cit. on p. 10).

Hincks, Rebecca (2002). "Speech synthesis for teaching lexical stress". In: *Proceedings of Fonetik, TMH-QPSR* 44.1, pp. 153–156 (cit. on p. 4).

Hincks, Rebecca and Jens Edlund (2009). "Promoting increased pitch variation in oral presentations with transient visual feedback". In: *Language Learning & Technology* 13.3, pp. 32–50 (cit. on p. 4).

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).

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. 5, 10).

Magen, Harriet S (1998). "The perception of foreign-accented speech". In: *Journal of Phonetics* 26.4, pp. 381–400 (cit. on p. 5).

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. 3, 6, 7).

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. 3).

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. 3).

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. 3, 7).

Peperkamp, Sharon and Emmanuel Dupoux (2002). "A typological study of stress 'deafness'". In: *Laboratory phonology* (cit. on pp. 4, 5).

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. 3, 8).

Project-Team PAROLE (2013). *Activity Report 2013*. Tech. rep. Nancy: LORIA (cit. on p. 3).

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. 9).

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. 5, 10).

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. 9).

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 p. 3).

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 p. 1).

Warren, Paul, Irina Elgort, and David Crabbe (2009). "Comprehensibility and prosody ratings for pronunciation software development". In: *Language Learning & Technology* 13.3, pp. 87–102 (cit. on p. 5).

Weber, Frederick and Kalika Bali (2010). "Enhancing ESL education in India with a reading tutor that listens". In: *Proceedings of the First ACM Symposium on Computing for Development - ACM DEV '10*. New York, New York, USA: ACM Press, p. 1 (cit. on p. 3).

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, 4).