PERCEPTUAL IMPACT OF SPEECH MELODY HYBRIDIZATION: ENGLISH AND CZECH ENGLISH*

JAN VOLÍN
Univerzita Karlova v Praze
jan.volin@ff.cuni.cz

KRISTÝNA POESOVÁ
Univerzita Karlova v Praze
kristyna.poesova@pedf.cuni.cz

Abstract
The current paper examines the role of intonation in the perception of foreign-accented speech. In order to assess how difficult it is to mentally process native, non-native and modified speech melodies, four conditions were analyzed and compared: native English, native English with Czech melody, Czech English with native melody and Czech English. The method of reaction times measurement in a word monitoring task was employed, in which 108 Czech listeners heard English sentences in the explored conditions and pressed a button when hearing a target word. Speech melody turned out to have a relatively weak but discernible impact on perceptual processing. Interestingly, Czech English proved to be more difficult to process than native English, although the listeners were Czech. The implementation of English F0 contours on Czech English speech slightly alleviated the cognitive load, however, the second hybrid, native English with Czech melody, pointed to the opposite direction. The causes of this discrepancy were investigated, particularly higher degrees of collocability in certain expressions.

Keywords: intonation, F0 contours, speech melody, reaction time, Czech-accented English

1. Introduction

Speech melody plays a number of important roles in language communication. Intonation systems characterized by a lexical tone distinction represent an unfamiliar concept for many European speakers even though the vast majority of Asian, African and indigenous American languages employ tones contrastively and partial tonal elements, typically attached to stressed positions, also occur in Europe, for example in Norwegian, Swedish, Croatian or Serbian (Collins and Mees 2013). Apart from Scandinavia and the Balkans, the interplay of tone and intonation has been investigated in detail in Limburgian dialects found mainly in

* This study was supported by the grant provided by the Czech Science Foundation, GAČR, under no. 14-08084S. The authors would like to thank all the volunteers who underwent reaction times measurements.

the area of south-eastern Netherlands and north-eastern Belgium (Gussenhoven and van der Vliet 1999; Gussenhoven 2004; Peters 2007).

In English and other European languages pitch and/or tones are not exploited for lexical purposes, rather, their variation helps to contextualize the uttered content by revealing the relationships between individual utterances and interlocutors. In addition, intonation adds various meanings to what is already expressed on the word level (Collins and Mees 2013). Not only can it support or expand the given meaning, it can frequently contradict it as in the case of irony. Furthermore, intonation has the power to bring certain parts of utterances in or out of focus, indicate major/foreground as opposed to minor/background information, signal turn continuation or completion, colour the words with numerous emotions, attitudes and interpersonal stances, group words coherently together or resolve syntactic ambiguities (e.g., Wells 2006; Wichmann et al. 2009). Although intonation is believed to be difficult to grasp at the conscious analytical level, L2 users of English, who often consider the function of intonation as purely decorative, should be informed about the various roles it plays in effective communication. Most importantly, they should learn to think continually about their listeners and send them the right prosodic signals to enable them to follow the intended message with ease and genuine involvement (Gilbert 2015).

Interestingly, the concern for listeners in the process of successful communication was reflected in Abercrombie's definition of a pronunciation goal more than sixty years ago. He cast doubt on achieving perfection in the area of pronunciation learning and teaching and proposed a more realistic aim for the majority of L2 users – *comfortable intelligibility*, which has been firmly advocated by pronunciation experts in the new millennium (e.g., Grant 2014; Derwing and Munro 2015). Abercrombie's understanding of *comfortable* refers to the pronunciation "which can be understood with little or no conscious effort on the part of the listener" (Abercrombie 1956: 37). Half a century later this idea resonated in one of Munro and Derwing's research-based dimensions of non-nativeness, *comprehensibility*, described as "listeners' perceptions of difficulty in understanding" (1995: 291). Evidence indicates that accented speech, though objectively intelligible, may receive lower comprehensibility scores due to increased processing difficulty. The key research questions that we address in the current paper is, to what extent individual cues of foreignness disrupt the smooth flow of perceptual processing, specifically in the area of speech melody.

When examining cognitive load it is crucial to realize that the extra burden experienced on the part of the listener may become discouraging, as mismatches between the expected forms and the incoming acoustic signal are likely to activate additional cognitive resources, which may take its toll in areas such as attention or working memory (Van Engen and Peelle 2014). Consequently, listeners may lose interest in productions that require too much effort and may even start avoiding future interactions or form negative evaluations of the speaker (Lippi-Green 1997). As many psycholinguistic experiments illustrate,

accented speech takes longer to be processed and information tends to be less accurate. Nevertheless, under the right circumstances, for instance with sufficient exposure to the target foreign accent, adaptation may take place relatively fast (for an overview see Cristia et al. 2012). Cognitive load can be made more tangible by measuring reaction times, which is a widely applied method in psycholinguistics, predominantly in spoken word recognitions tasks (Grosjean and Frauenfelder 1996). Faster reaction times to a given stimulus correlate with smoother processing whereas slower latencies indicate more demanding processing of the signal. Racine (2013) warns about possible intricacy of these measurements and gives useful recommendations to those wishing to exploit this format to its fullest.

Returning to the desirable outcome of foreign language learning and teaching, *comfortable intelligibility*, one cannot avoid raising a crucial question: to what extent do individual acoustic features or their combinations make the speech sound foreign and thus less comprehensible? The prosodic research has shown that not only segmental but also suprasegmental features contribute to perceived accentedness (Anderson-Hsieh, Johnson and Koehler 1992; Derwing and Munro 1998; Derwing and Rossiter 2003; Field 2005). For instance, in a study aimed at describing the general relevance of prosody to the perception of a foreign accent Jilka (2000) confirmed its significance in a series of experiments using low-pass filtered stimuli on German-accented English and English-accented German. In addition, his data suggested that intonation was a stronger indicator of a foreign accent than other prosodic aspects (rhythm and speaking rate), however weaker than segmental features. Hahn (2004) opted for the nucleus placement as her main research variable and found out that its correct location facilitated comprehension as opposed to its misplacement or total absence. Kang et al.'s (2010) exhaustive research established 29 prosodic aspects of accentedness comprising rate, pause, stress and pitch measures and investigated their relationship with L2 comprehensibility and proficiency assessments. Rather than using impressionistic judgements, a computer-based acoustic analysis, which yielded more objective results, was employed. The parameters accounted collectively for 50% of variance in the ratings, and the findings also corroborated a substantial contribution of suprasegmental errors to perceived accentedness.

Looking at intonation research carried out on Czech-accented English, the majority of studies seem to be production-oriented. Volín et al. (2015) analysed F0 tracks in the speech of professional newsreaders from the BBC and the Czech National Radio employing distributional measures (Volín and Bartůňková 2015) in order to establish the reference values for pitch level, pitch span and downtrend gradient in English and Czech respectively. Interestingly, in the subsequent measurement of Czech-accented English, the interference hypothesis, according to which the interlanguage values should lie between the two natural languages, was not supported. There was a clear difference in pitch span between the Czech and English, but the pitch span of Czech-accented

English was the narrowest. Interestingly, in a study with a similar research design but different speech material, English-accented Czech, the authors revealed some striking similarities with the previous findings – the accented productions do not appear to be mere mixtures or compromises between the mother tongue and the target language (Galeone et al. 2015).

As none of the previously mentioned studies provided perceptual validation of Czech-English melodies, the current work focuses solely on the perception domain. More specifically, we are interested in what impact modified F0 contours will exert on perceptual processing. As reaction time measurements proved to be capable of capturing the intricacies of cerebral processing, this methodology was selected for investigating the cognitive effects of manipulated speech. While the zero hypothesis states no changes in reaction latencies, the alternative hypothesis presupposes smoother processing in case of nativelike alterations on the one hand and, on the other hand, an increase of listening effort in case of adding foreign elements to native F0 contours.

2. Method

Two native speakers of standard British English and two intermediate Czech speakers of English with an estimated medium-scale foreign accent recorded a set of eight semantically unpredictable sequences, e.g., *Mutual admiration after so many meetings without* ***street*** *or backyard plotting* or *The previous longitude* ***note*** *directly changes their relationship*. Semantic unpredictability is required in reaction time (RT) paradigms, since ordinary texts contain collocations that would put a word in the listeners' minds even before this word is actually spoken. The utterances contained target words in various positions relative to their beginning. In the examples above, the target words are in bold. Again, the RT experiments avoid identical location of the tested items as it might lead to unconscious adaptation and expectations that would bias the reactions to the target words. (For the whole set see Appendix.)

Pairs of stimuli were created such that one member of a pair was the original recording, the other received an intonation contour from a counterpart speaker. The contours were implanted by PSOLA re-synthesis algorithm (Boersma and Weenink 2014). To preserve naturalness of the utterances, adjustments had to be made for the nuclei positions and for the overall pitch level. This means that regardless of possible different timing, the transplanted F0 contour had peaks and valleys on the same syllables as the original utterance. Also, the whole contour sometimes needed to be shifted up or down to match the register of the speaker on whose segments it was transplanted.

The experimental design then comprised four conditions: (1) Eng-Eng – native English speech with native English intonation; (2) CzE-CzE – Czech English speech with Czech English intonation; (3) Eng-CzE – Native English speech with Czech English intonation; (4) CzE-Eng – Czech English speech

with native English intonation. Conditions 1 and 2 are original recordings (resynthesized without changes in order to equalize the technical quality of the sound), while conditions 3 and 4 are hybridized. Naturally, we could not use one sentence for all four conditions, since the listeners must not be familiar with the wordings if we want them to react to a target.

The items were randomly mixed into large sets of fillers (distractors) and split into four blocks to avoid listeners' fatigue. The fillers contained speech material for other research purposes. The test session took approximately 20 minutes including the trial run. The four blocks were separated with three short breaks during which a picture appeared on the screen and the participants held short conversations with the experimenter. The test was administered in four differently randomized testing loops.

There were 108 participants who were asked to listen to English sentences and press a button as quickly as possible when hearing a target word (i.e., the word monitoring paradigm). The respondents were Czech university undergraduates studying the English language either as one of their majors or as a specialization. Their language level varied between intermediate and upper-intermediate. The perceptual experiment was performed in a sound-proof booth at the Institute of Phonetics in Prague using the DMDX display software (Forster and Forster 2003) and the Black Box ToolKit. The participants were instructed to leave two fingers on the special button and press it as quickly as possible once they identified the target word in the stimulus. The data collection took slightly over one month.

3. Results

As reaction time is a delicate measure, the raw output of RT experiments has to be pre-processed before the results can be calculated. It is necessary to discard the reactions that are below and over certain thresholds. Some responses are too fast to count as thinkable human reactions to a stimulus: certain speeds are just not neurophysiologically possible. Conversely, after an unusually long time a response is assumed to be not a simple reaction but a result of considerations or hesitations. In our study we opted for commonly found values of 150 milliseconds for the lower threshold and 1200 milliseconds for the upper threshold.

The data prepared as described above were submitted to one-way ANOVA with the condition (see Method) as the independent variable and reaction times as the dependent variable. The algorithm returned a significant result for the main effect of condition: $F(3, 1406) = 2.98$; $p = 0.03$. The situation is captured in Figure 1. It can be observed that the longest reactions are connected with Czech English with no manipulations, i.e., with both segmental and melodic features as they were uttered by the Czech learners of English as a foreign language. When these sentences were given with F0 tracks taken from native

English production, the mean reaction time dropped by 13 ms. The reaction to native English (Eng-Eng) is 30 ms faster. Quite unexpectedly, native English speech with the Czech F0 track produced the shortest reaction times (although the difference here is only 8 ms and only after data cleansing – with raw data Eng-Eng was faster than Eng-Cz).

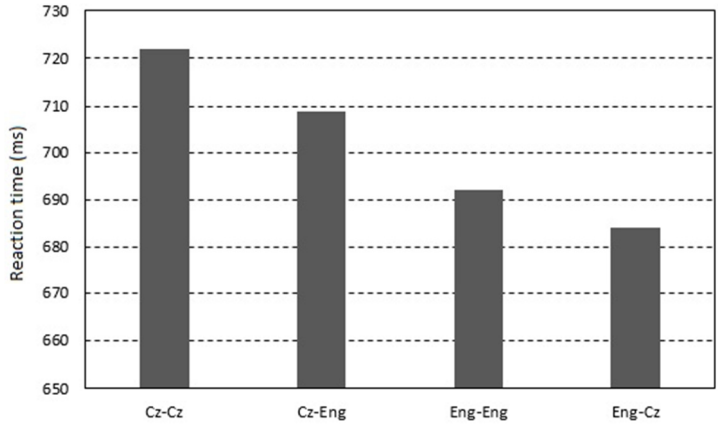

Figure 1. Mean reaction times to English sentences produced by Czech learners of English (Cz-Cz), native speakers of English (Eng-Eng), and hybrid versions with interchanged pitch contours (Cz-Eng = Czech speaker with English intonation; Eng-Cz = English speaker with Czech intonation).

The post-hoc Tukey test found a significant difference between the Cz-Cz and Eng-Cz conditions and marginally significant difference between Cz-Cz and Eng-Eng conditions. However, significance of these results improved when some of the items were excluded from analyses. This will be reported below. Before that we would like to present some further general results related to the sample as a whole.

One of the indicators of processing difficulty apart from the reaction time itself is the number of items that did not receive any reaction or where reaction times exceeded the upper threshold. Table 1 presents the counts found in our data set. It is clear again that Czech-accented English with its original melody (Cze-Cze) causes most problems, while native English is the easiest to process. Implanting English melody on Czech speech (Cze-Eng) slightly alleviated the processing burden and implanting Czech melody on English speech produced the opposite result. The statistical significance of the differences was confirmed by a chi-square test: χ^2 (3) = 10.24; $p < 0.05$. It should be noted, however, that the greatest contributor to the significance is the difference between the second and the third columns (CzE-Eng vs. Eng-CzE).

Table 1. Numbers of failed responses in the data set under the four test conditions. No response = the subject failed to react completely; long response = RT higher than 1200 ms.

	CzE-CzE	*CzE-Eng*	*Eng-CzE*	*Eng-Eng*
No response	26	22	9	7
Long response	68	66	55	53
Both problems	94	83	64	60

As hinted above, the individual sentences in our set did not behave uniformly. Figure 2 displays the differences in reaction times between CzE-CzE and CzE-Eng conditions for our male and female Czech speakers. Apart from the obvious slower reactions to the female speaker, it can be observed that for one of the sentences (M1 and F1), there is a difference over 30 ms between the conditions, while the other two sentences (M2 and F2) produced only 2 ms and 12 ms respectively. This means that the same speakers and analogous manipulation (i.e., implantation of the native English melody on the Czech-accented speech) led to different effects.

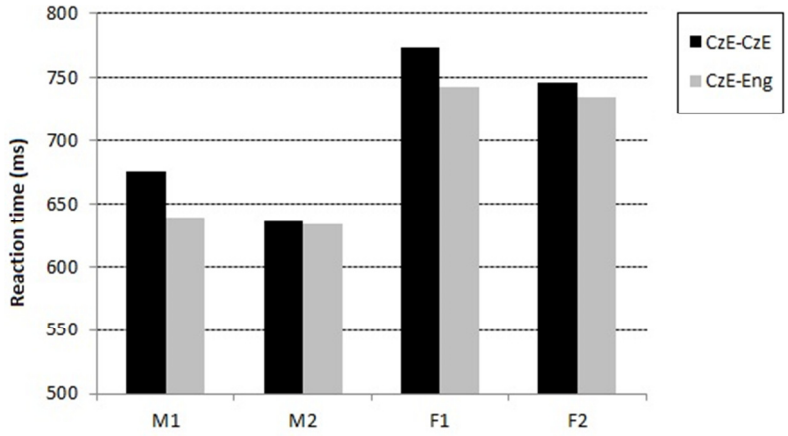

Figure 2. Reaction time differences between items produced by Czech learners of English (black columns CzE-CzE), and the same sentences with native English intonation (grey columns CzE-Eng) implanted in them. Sentences M1 and M2 were produced by a male Czech, F1 and F2 by a female Czech speaker.

The two main causes of this discrepancy could be (a) the specific difference between the implanted and original contour, and (b) some sort of hidden semantic cue. In the following paragraphs we will consider both. Although we took care to produce sequences without semantic predictability (see Appendix), some collocability was found under scrutiny with corpus tools. The *Araneum Anglicum Maius Corpus* (Benko 2014) with 1,200,048,075 tokens was used to

calculate the logDice measure of collocation of the target word with the three preceding words. The logDice scores are most commonly numbers between 0 and 10, although lower and higher values are occasionally possible. The score for M1 was 3.92, whereas for M2 it was 5.05. Similarly, the target in F1 produced 2.77, whereas F2 reached 4.66. Clearly, the items with higher collocability were also items with lower difference in reaction times and *vice versa*. Although this looks quite indicative, further research would be needed to confirm a causal relationship.

To check the magnitude of differences between the original and manipulated items, we overlaid the paired F0 tracks into one plot. They can be inspected in Figure 3 and 4. Again, it can be observed that there are greater differences between the tracks in M1 than in M2 and, analogically between the tracks in F1 and F2 (consider the space between the F0 tracks in each plot). The items with greater difference between the F0 tracks also produced greater differences in reaction times (see Figure 2).

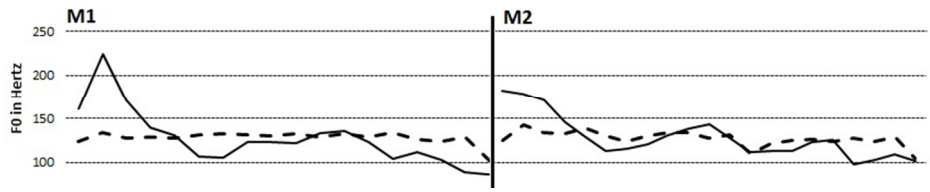

Figure 3. Comparison of the original and implanted F0 tracks in items M1 and M2 (see text). The solid line is the native English F0 track, the dashed line is the Czech English rendering.

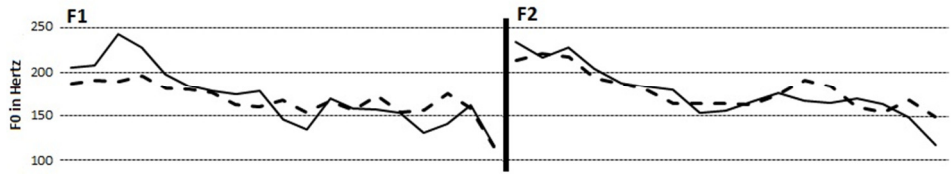

Figure 4. Comparison of the original and implanted F0 tracks in items M1 and M2 (see text). The solid line is the native English F0 track, the dashed line is the Czech English rendering.

4. Discussion

We infer from the current results that melody of speech is a relevant component of foreign accent and, importantly, it contributes to the cognitive load in perception of foreign-accented speech. Compared with other elements in the speech signal, however, its impact is not strong. Reactions to native English (Eng-Eng) sentences were virtually the same as reactions to native English speech with Czech melodies implanted on it. On the other hand, apart from the

ANOVA results, the existence of differences was also observed in the number of items with no or extra-long responses (Table 1).

Although the listeners in our experiments were Czech learners of English, Czech accented English was the most difficult for them to process mentally. Native English was the least demanding. Despite the common belief that foreign-accented speakers understand each other better than they understand native speakers, our respondents reacted best to the native English sentences. It is not the case then that speakers of a certain variant of English find their own variant the easiest to understand. This could be perhaps explained through the effect of exposure. Our Czech learners of English were university students who, by this time of acquisition, are most probably exposed more to American or British accents than to Czech-accented English. This hypothesis could be tested with schoolchildren who will have less experience with American films and British pop-music than with the speech of their Czech teachers or schoolmates from English classes.

Since to the best of our knowledge there is no research linking collocation scores to reaction times in any straightforward manner, we are unable to say, to what extent logDice measures can be held responsible for the differences in the results for individual sentences. Nevertheless, they point in the right direction: a greater association score was found for shorter reaction times.

In the very near future we intend to test the influence of the magnitude of F0 track differences on the reaction times. Figures 3 and 4 lead us to a question whether it is the initial alerting peak found in native English contours or whether it is the global sum of differences (or perhaps some other factors). The trouble with these questions is that they might require more artificial conditions, which is exactly what we would like to avoid. Our F0 tracks were real trajectories of fundamental frequency found in read texts. One of the most difficult tasks for the follow-up research will be finding out how to test melodic effects without creating unrealistic contours.

Another appealing idea is to test listeners of other linguistic experience. It would be quite interesting to know whether native English respondents or listeners who are unfamiliar with Czech accented English (e.g., Finnish or Portuguese) would produce similar results to ours. To what extent can we count on the existence of salient (maybe universal) perceptual features of foreignness (see, e.g., Major 2007) against the effects of exposure (see above)? We believe that future experimenting will bring noteworthy answers as well as the inevitable further questions.

References

Abercrombie, D. 1956. *Problems and Principles. Studies in the Teaching of English as a Second Language*. London: Longmans, Green.

Anderson-Hsieh, J., Johnson, R. and K. Koehler. 1992. The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. Language Learning 42, 529–555.

Boersma, P. and D. Weenink. 2014. *Praat: doing phonetics by computer*. Version 5.4.06. [Computer programme]. Available from: http://www.praat.org/.

Collins, B. and I. M. Mees. 2013. *Practical Phonetics and Phonology*. Abingdon: Routledge.

Cristie, A., Seidl, A., Vaughn, Ch., Schmale, R., Bradlow, A. and C. Floccia. 2012. Linguistic processing of accented speech across the lifespan. Frontiers in Psychology 3, 1–15.

Derwing, T., Munro, M. J. and G. Wiebe. 1998. Evidence in favor of a broad framework for pronunciation instruction. Language Learning 48, 393–410.

Derwing, T. M. and M. J. Rossiter. 2003. The Effects of Pronunciation Instruction on the Accuracy, Fluency, and Complexity of L2 Accented Speech. Applied Language Learning 13, 1–17.

Field, J. 2005. Intelligibility and the Listener: The Role of Lexical Stress. TESOL Quarterly 39 (3), 399–423.

Forster, K. I. and J. C. Forster. 2003. DMDX: A Windows Display Program with Millisecond Accuracy. Behavior Research Methods, Instruments, and Computers 35, 116–124.

Galeone, D., Johnson, W. And J. Volín. 2015. Intonation contours in English Czech and Czech English. In Adamczyk, M. (ed.), *Accents 2015. The Book of Abstracts*, 12. Lodź: University of Lodź.

Gilbert, J. 2014. Myth 4: Intonation is hard to teach. In J. Levis (ed.), *Pronunciation Myths: Applying Second Language Research to Classroom Teaching*, 107-136. Ann Arbor, MI: University of Michigan Press.

Grant, L. 2014. *Pronunciation Myths. Applying Second Language Research to Classroom Teaching*. Ann Arbour: University of Michigan Press.

Grosjean, F. and U. H. Frauenfelder. 1996. A Guide to Spoken Word Recognition Paradigms: Introduction. Language and Cognitive Processes, 11(6), 553–558.

Gussenhoven, C. and P. van der Vliet. 1999. The phonology of tone and intonation in the Dutch dialect of Venlo. Journal of Linguistics 35, 99–135.

Gussenhoven, C. 2004. Tone in Germanic: Comparing Limburgian with Swedish. In G. Fant, Fujisaki, H., Cao, J. and Y. Xu (eds.), *From traditional phonology to modern speech processing: Festschrift for Professor Wu Zongji's 95th birthday,* 129–136. Beijing: Foreign Language Teaching and Research.

Hahn, L. D. 2004. Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals. TESOL Quarterly 38(2), 201–223.

Jilka, M. 2000. Testing the contribution of prosody to the perception of foreign accent. Dissertationsschrift zur Dr. phil, Fakultät für Philosophie der Universtität Stuttgart.

Kang, O., Rubin, D. and L. Pickering. 2010. Suprasegmental Measures of Accentedness and Judgments of Language Learner Proficiency in Oral English. The Modern Language Journal 94(4), 554 – 566.

Keating, P. and G. Kuo. 2010. Comparison of speaking fundamental frequency in English and Mandarin. UCLA Working Papers in Phonetics 108, 164–187. Los Angeles: University of California.

Lippi-Green, R. 2012. *English with an accent: language, ideology and discrimination in the United States*. London, UK: Routledge.

Major, R. C. 2007. Identifying a Foreign Accent in an Unfamiliar Language. Studies in Second Language Acquisition 29, 539–556.

Munro, M. J. and T. M. Derwing. 1995. Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. Language and Speech 38 (3), 289–306.
Munro, M. J. and T. M. Derwing. 2005. Second language accent and pronunciation teaching: A research based approach. TESOL Quarterly 39(3), 379–397.
Munro, M. J. and T. M. Derwing. 2015. *Pronunciation Fundamentals*. Amsterdam: John Benjamins Publishing Company.
Peters, J. 2007. Tone and Quantity in the Limburgian Dialect of Neerpelt. In J. Trouvain and W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences: Saarbrücken, Germany*, 1265-1268. Universität des Saarlandes.
Racine, J. P. 2013. Reaction Time Methodologies and Lexical Access in Applied Linguistics. Vocabulary Learning and Instruction.
Available from http://vli-journal.org/issues/onlinefirst/vli.v03.1.racine.pdf [Accessed: 10 November 2015].
Reed, M. and T. Jones. 2015. The Melody of English: Research and Resources for Teaching the Pragmatic Functions of Intonation. IATEFL PronSIG webinar held 17th February 2015.
Rogerson-Revell, P. 2011. *English Phonology and Pronunciation Teaching*. London: Continuum.
Van Engen, K. J. and J. E. Peelle. 2014. Listening Effort and Accented Speech. Frontiers In Human Neuroscience 8, 1–4.
Volín, J., Poesová, K. and L. Weingartová. 2015. Speech Melody Properties in English, Czech and Czech English: Reference and Inteference. Research in Language 13(1), 107–123.
Volín, J. and H. Bartůňková. 2015. Assets and Liabilities of Simple Descriptors of Fundamental Frequency Tracks. In O. Niebuhr and R. Skarnitzl (eds.), *Tackling the Complexity in Speech,* 147-161. Prague: Charles University.
Wells, J. C. 2006. *English Intonation*. Cambridge: Cambridge University Press.
Wichmann, A., Dehé, N. and D. Barth-Weingarten. 2009. Where prosody meets pragmatics: research at the interface. In D. Barth-Weingarten, N. Dehé and A. Wichmann (eds.), *Where prosody meets pragmatics,* 1-20. Bingley: Emerald.

Appendix

The set of semantically unpredictable sentence used in the experiment. The target words are in bold letters.

*1. Mutual admiration after so many meetings without **street** or backyard plotting*
*2. The previous longitude **note** directly changes their relationship.*
*3. Several light yellow skyscrapers with **down** and up pointing poles*
*4. Writing predictable reports gives them **step** and jump strategies*
*5. A colourful jacket and brownish belts all over **plates** made of plastic*
*6. An impressive greyish horizon **tower** provided an easy target*
*7. German administrators opened the earlier **teach** and learn project*
*8. Eastern fighters **deal** with their opponents with decency and grace*