FILOZOFICKÁ FAKULTA
UNIVERZITY KARLOVY
V PRAZE

ÚSTAV ANGLICKÉHO JAZYKA
A DIDAKTIKY

Diplomová práce

Bc. Jakub Kaprál

Rhythm sensitivity to speech and non-speech stimuli in musically trained and untrained population

Rytmická senzitivita hudobne školeného a neškoleného človeka voči rečovým a nerečovým podnetom

PRAHA 2017

Vedoucí práce:
doc. PhDr. Jan Volín, Ph.D.

Acknowledgments

I would like to thank my thesis supervisor doc. PhDr. Jan Volín, Ph.D. for his guidance, invaluable comments and advice, for his willingness to be available for consultations both in person and through e-mails, and for his indispensable assistance with the experiment planning. I would also like to thank all the people who participated in the experiment, and above all I would like to thank my loving parents who have never ceased to support me in my university studies both verbally and materially.

I declare that the following MA thesis is my own work for which I used only the sources and literature mentioned, and that this thesis has not been used in the course of other university studies or in order to acquire the same or another type of diploma.

Prehlasujem, že som diplomovú prácu vypracoval samostatne, že som riadne citoval všetky použité pramene a literatúru, a že práca nebola využitá v rámci iného vysokoškolského štúdia alebo k získaniu iného alebo rovnakého titulu.

V Prahe, 8.8.2017 ….....

Jakub Kaprál

Abstract

The purpose of this diploma thesis is to analyse the ability of the human ear to hear slight rhythm deviations in speech and non-speech phrases. The first part contains theoretical background for the study of speech rhythm and summarizes the research that has been already conducted in this area. It focuses especially on the perceptual nature of rhythm, the concept of P-centers, and provides a comparative study of speech rhythm and musical rhythm and their common properties and functions. The theoretical part is concluded with the analysis of potential influences of linguistic and musical training on the production and perception of rhythm, and hypotheses and research questions are formulated.

The practical part contains a perceptual experiment designed to examine the ability to identify rhythm manipulations in short speech and non-speech, i.e. percussive, phrases. Short English phrases are selected and their rhythmically altered counterparts are prepared. Participants are then presented with pairs of speech or non-speech phrases and a task to identify rhythmical discrepancies between them.

The results highlighted several differences between the nature of speech and non-speech rhythm. While the presence of stressed syllables enhances perception of rhythm deviations in speech, this is not the case for the non-speech signal. Performance in the experiment deteriorates with the increasing length of phrases, and rhythm manipulations become identifiable when their duration reaches approximately 60-70 milliseconds. However, no correlations between past or present musical training and the performance in the experiment was found. The only personal variable which influenced the results and improved performance in the experiment was the participants' level of L2 English. Weak performance in the cross-modal test, which followed assumed positions of P-centers, gives rise to questions about the influence of other factors on its precise location within the phrase.

Keywords: perception, rhythm sensitivity, P-center, musical training

Abstrakt

Cieľom tejto diplomovej práce je analyzovať schopnosť ľudského ucha počuť malé rytmické odchýlky v rečových a nerečových frázach. Prvá časť obsahuje teoretický základ pre štúdium rečového rytmu a sumarizuje výskum, ktorý bol doposiaľ v tejto oblasti vykonaný. Zameriava sa predovšetkým na percepčnú povahu rytmu, koncept P-centier, a prináša komparatívnu štúdiu rečového a hudobného rytmu s ich spoločnými vlastnosťami a funkciami. Teoretická časť je zakončená analýzou potenciálnych vplyvov lingvistického a hudobného tréningu na produkciu a percepciu rytmu, a následne sú sformulované hypotézy a výskumné otázky.

Praktická časť obsahuje percepčný experiment, ktorý bol vytvorený za účelom preskúmania schopnosti identifikovať rytmické manipulácie v krátkych rečových a nerečových, tj. perkusívnych, frázach. Na experiment sme vybrali krátke anglické frázy a ich rytmicky pozmenené ekvivalenty. Respondentom boli následne predstavené páry rečových a nerečových fráz a ich úlohou bolo identifikovať rytmické nezrovnalosti medzi nimi.

Výsledky zvýraznili niekoľko rozdielov v povahe rečového a nerečového rytmu. Kým prítomnosť prízvučných slabík napomáha vnímaniu rytmických odchýlok v reči, pre nerečový signál takáto skutočnosť preukázaná nebola. Úspešnosť v experimente sa znižuje spolu s narastajúcou dĺžkou fráz, a rytmické manipulácie sú identifikovateľné keď ich dĺžka dosahuje približne 60-70 milisekúnd. Žiadna korelácia však nebola preukázaná medzi súčasným alebo predchádzajúcim hudobným tréningom a úspešnosťou v experimente. Jediná premenná, ktorá ovplyvnila výsledky a zlepšila úspešnosť v experimente bola úroveň anglického jazyka. Nízka úspešnosť v cross-modálnom bloku, ktorý využíval predpokladané umiestnenie P-centier v reči, predkladá otázky o vplyve iných faktorov na ich presnú polohu v rámci rečovej fráze.

Kľúčové slová: percepcia, rytmická senzitivita, P-centrum, hudobný tréning

Table of contents

1. Introduction

There have been numerous studies investigating speech rhythm from the perceptual point of view, and many phoneticians and psychologists are suggesting the existence of so-called perceptual centres occurring in speech. These are perceptually relevant moments occurring in the brain grouped into more or less regular patterns that aid the listener in the perception of speech rhythm.

This diploma thesis will examine the perceptual nature of rhythm in speech and music. The theoretical part of the present study contains theoretical background concerned with the concept of rhythm and its realization in speech. Speech rhythm theory is summarized in the section 2.1. The first two sections 2.1.1 and 2.1.2 summarize views on language rhythm in general and on its prominent features in the English language. Sections 2.1.3 and 2.1.4 contain an analysis of English stress and its acoustic correlates. The following sections 2.1.6 and 2.1.7 contain an overview of the rhythm metrics research and its limitations.

The section 2.2 focuses on the perceptual nature of speech rhythm. Research conducted on early language discrimination and its results are presented in the section 2.2.1. An overview of P-center research is provided in the section 2.2.2 as well as criticism and problematic issues related to its precise location. Techniques used for the resynthesis of speech are summarized in the section 2.2.3.

The relationship between music and speech is analysed in the section 2.3. Theories of common origin are presented in the section 2.3.1, and the shared properties and functions of music and speech are analysed in the section 2.3.2. Section 2.3.3 describes cognitive processing mechanisms which are used for music and speech. The section 2.3.4 focuses on the perception of musical rhythm and its most prominent features, followed by the section 2.3.5 collecting evidence from neurophysiology and related fields. The last section 2.3.6. analyses potential influence that musical and linguistic training can have on the production or perception of rhythm.

The practical part contains a perceptual experiment studying sensitivity of the human ear to rhythmic deviations. Chapter 3 describes the material which was used for the perceptual task and the method which governed the experiment. Section 3.1 contains information on recordings that were chosen for the experiment and the temporal modifications that were applied to them.

The reasons for the selection of participants are explained in the section 3.2. The experiment, along with the introductory training, is described in the section 3.3.

Results of our experiment are presented in the chapter 4. Sections 4.1–4.6 analyse the results influenced by various variables, such as the ordering of the experiment, the presence of temporal manipulation in the phrase, the scale of manipulation, the presence of stress in the target element, the length of the phrase, or the position of the modification within the phrase. Intra-speaker consistency is evaluated in the section 4.7. The results for individual stimuli are analysed in the section 4.8. The section 4.9 connects the results to the length of musical training, intensity of practice and command of English.

Chapter 5 contains general discussion which evaluates our results in detail, assesses the hypotheses and provides answers to the research questions. It also attempts to explain inconsistencies in the results of the experiment. Chapter 6 presents a conclusion which summarizes the findings of the present study.

2. Theoretical background

The beauty of communication lies in its complexity. The possibilities how information can travel from the agent to the recipient are manifold; people can express themselves through their gestures, movement, or voice. One can say that even silence speaks, but what is professed to be the feature that distinguishes human species from other animals is speech. Philosophers such as Rousseau, Darwin and Spencer articulated theories of common origin of language and music which suggested that the first prehistoric languages may have been sung, not spoken:

> Musical notes and rhythm were first acquired by the male or female progenitors of mankind for the sake of charming the opposite sex. (Darwin, 1871/1981, p. 336)

Even though being only hypotheses, the theories favour close relationship between music and language. Both are tools for conveying, to various degree, emotion and information. Music and language share common properties, such as rhythm and melody. These are essential for the organisation of both of them. Rhythm describes the temporal organisation of speech and music, whereas melody controls the pitch of an utterance or musical motif. Each of them is indispensable for effective communication. It was found out that sequences composed of events organized around a regular beat – also called rhythm – are easier to perceive and reproduce than are sequences without such organisation (Povel & Essens, 1985). Melody, termed intonation in the linguistic realm, is necessary for signalling the type of utterance. If the only cues provided were stress and grammatical means such as word order, it would be difficult to say whether the speaker produced a statement, question or order. Speech without the proper use of intonation seems dull and the recipient can easily lose concentration.

2.1 Speech rhythm theory

Rhythm is one of the underlying organising principles in nature, be it seasonal changes, breathing, movement or heartbeat of any animate being. All art forms – for example dance, music, poetry, or architecture – contain rhythmical patterns expressed in tactile, auditory or visual terms. This is related to the fact that rhythm is subject to multisensory perception. Contrast creates rhythm. In nature it can be the alternation between day and night, movement and stillness, contraction and release, silence and sound etc. In speech, rhythm is created by

alteration of stressed and unstressed syllables, and to a certain degree by alternation of consonants and vowels.

2.1.1 Language rhythm

Speech rhythm has been observed to have practical implications, and speech rhythm is largely cultivated in the art of rhetoric, as Cicero noted in *On the Ideal Orator*:

> In all sounds and utterances rhythm is understood as the quality of having certain beats and of being measurable by regular intervals […] if we are right to think that a constant steady flow of babble without pauses is crude and unpolished, the reason for this rejection is surely that it is natural for the human ear to measure the rhythm of the sounds that are produced by a voice, and that this is impossible if they don't have any rhythm; rhythm is the product of separation, of a beat at regular, or often varying intervals. We can discern it in falling drops of water (because they are separated by intervals), but not in an onrushing stream.

Cicero saw that what distinguishes a skilled orator from the others is his mastery over rhythm:

> Among the many things that distinguish the orator from those unskilled and inexperienced in speaking, there is nothing that does so more than this: the unschooled speaker crudely pours out as much as he can, and lets his breath, not art, determine the limits of what he says. The orator, on the other hand, so ties his thoughts to the words, that all of them are encompassed by a kind of rhythm that is at once confined and free. For after fastening the thoughts in the bonds of form and cadence, he loosens and frees them by changing the order, so that the words are neither confined as if by some fixed law of verse, nor so free that they just wander about.

Rhythmical organisation of speech is related to the cognitive abilities of humans. Neurophysiologists showed that neuronal signals are emitted from the brain in the form of short electrical pulses instead of a constant flow of information. A fluent rhythmical patterning is therefore more easily processed by the listener's brain as he can 'tune in' to the rhythm of the speaker. Dysfluencies result in a more demanding process of cognition requiring more energy for the listener to process the information. Perceptual experiments of Ghitza and Greenberg

(2009) provided some evidence for the claim that speech perception is simpler for the listener when the incoming information contains predictable rhythm.

2.1.2 English language rhythm

There has been a considerable body of research on language rhythm carried out in recent decades. It has been based on the idea of *isochrony*, assuming that speech units occur at roughly regular intervals. First accounts of language rhythm were impressionistic (Lloyd James, 1940), describing prototypical Germanic languages such as English and Dutch as having a 'morse-code rhythm' and prototypical Romance languages such as French and Spanish as having a 'machine-gun rhythm'. These terms are elucidated by Pike (1945), who coined the terms 'stress-timed' and 'syllable-timed' for the distinction between different language rhythms. This typological distinction was initially related to isochronous speech intervals, following the hypothesis that in syllable-timed languages the units recurring at equal intervals are syllables and in stress-timed languages the units are stress-delimited feet. In *Elements of General Phonetics*, Abercrombie (1967) considers this distinction to be clear-cut, dividing all world languages into either class:

> As far as is known, every language in the world is spoken with one kind of rhythm or with the other. In the one kind, known as a syllable-timed rhythm, the periodic recurrence of movement is supplied by the syllable-producing process [...] the syllables recur at equal intervals of time - they are isochronous. […] In the other kind, known as a stress-timed rhythm, the periodic recurrence of movement is supplied by the stress-producing process […] the stressed syllables are isochronous. (Abercrombie, 1967:97)

Although some phoneticians held this terminology to be categorical, the division was considered by Pike (1945) to be not strict but rather a continuum on which the languages occur, possessing more or less features of the one or other prototype. Later, the rhythmic classification was extended with the inclusion of 'mora-timed' languages such as Japanese or Slovak (see Warner and Arai, 2001 for an overview).

Lehiste (1973, 1977) examined the concept of isochrony in perception-focused experiments. The subjects were presented with both speech and non-speech stimuli. Whereas the perception of duration of non-speech material was accurate, the speech material was perceived to be more regular than it was in reality.

There have been several main differences pointed out between the stress-timed and syllable-timed languages that became the basis for the creation of rhythm metrics. In stress-timed languages, the degree of vowel reduction in unstressed syllables is much greater than in syllable-timed languages, which makes stressed syllables relatively salient (Dasher and Bolinger, 1982; Roach, 1982). Stress-timed languages also have a wide variety of syllable structures with greater complexity allowable in onsets and codas as opposed to more restricted choice in syllable-timed languages (Dauer, 1983, 1987). In stress-timed languages, the trend of heavier syllables attracting stress is much stronger than in syllable-timed languages. It was also found that open syllables (CV structures) are more widespread in syllable-timed languages.

2.1.3 English stress

According to Roach (1991), stress can be studied from the productional and perceptual point of view. Production of stress is claimed to be related to the muscular action of the speaker. The prominent syllable is marked by variations in four acoustic cues: fundamental frequency F0, amplitude, duration and formant structure. The perceptual correlates of these four acoustic cues are respectively: pitch, volume, length and a different timbre to the vowel (Frost, 2011).

The placement of stress in English is a complex matter. Its location can be predicted on the basis of several factors, such as complexity of the word, its grammatical category, number of syllables and their phonological structure, as only strong syllables can be stressed. English words often have variable stress caused by conversion – noun-verb, adjective-verb etc. – which usually moves the stress to the adjacent syllable.

Stress in English exists at different levels. It is a relative phenomenon, and majority of linguists agree that it is marked by several degrees of prominence. Individual words are marked by a fixed lexical stress, but there is a much more complex behaviour of stress above word level referred to as *sentence stress* which moves the focus according to the utterer's discursive intention (Frost, 2011). Sentence stress may be assigned to any syllable and it "gives prominence to the syllables that are lexically stressed, primarily by assigning them a pitch accent" (Xu & Xu 2005: 160). Besides lexical stress and sentence stress, any stressed syllable might additionally receive highly unpredictable contrastive stress, also known as 'narrow focus' (Ladd 2008: 216). These complexities explain that although English has a free word stress, its realization depends on the function of the sentence and intentions of the speaker.

In English language two levels of stress are widely recognized: primary and secondary stress

which is slightly weaker than the former (Roach, 1992). However, there have been differing views on how many stress levels exist in English. According to Pennington (1996: 131–132), four to six stress levels are sufficient for a detailed transcription. On the other hand, Cruttenden (1986:21) distinguishes only four stress levels: primary, secondary, tertiary stress and the unstressed syllable.

English speech rhythm is thus composed of a succession of more or less prominent syllables. In terms of isochrony, English has tendency to regulate its rhythm by the variously stressed syllables, interspersed with unstressed syllables.

2.1.4 Acoustic correlates of stress

As identified by Roach (1991), there are four acoustic correlates of stress in current research on prominence (as presented in Frost, 2011; Plag et al., 2011): pitch (F0), duration (quantity), intensity (loudness or amplitude) and vowel quality (formant structure). Tendency of English stressed syllables is to have higher pitch, longer duration, higher duration and the full range of vowel phonemes (Plag et al, 2011: 362). However, this is true for English on which the research was carried out. Other languages differ in the extent to which these correlates apply in their stressed syllables. In terms of duration, English stressed syllables are approximately 1,5 times longer than unstressed ones (Dauer, 1983). The difference in length is also supported by the fact that English employs an extensive vowel reduction in unstressed syllables.

Phonetic research was attempting since the middle of 20th century to find out which of the four correlates is the most important cue for the listener. First experiments by Fry (1955, 1958) on converted pairs with different stress pattern (e.g. '*permit* as a noun, *per'mit* as a verb) found F0 to be more reliable than duration and intensity. Several years later (1965), Fry included the aspect of formant structure as a cue for stress perception, but his results are questionable due to the limits caused by the technology available at the time. Bolinger (1958) repeated Fry's experiment and hypothesized a pitch accent theory assessing 'pitch prominence' as a chief cue of stress. Bolinger described it as "a rapid and relatively wide departure from a smooth or undulating contour". Bolinger also assumed that greater duration of stressed syllables is only a result of a pitch movement: "A pitch obtrusion requires time for its execution. When the pitch accent is embraced completely by a single syllable, the syllable is lengthened to accommodate the necessary range of pitches […] Figuratively speaking, it is there in order to make room for the accent" (Bolinger, 1958; pp. 138).

Although phoneticians vary in their opinion on the importance of cues for stress perception, pitch seems to be the strongest indicator for stress. In a production experiment, Lieberman (1960) found that F0 was a reliable cue in 90% of the cases, intensity (peak envelope) in 87%, and a duration in 66%. His findings agree with Morton & Jassem (1965) who found F0 to be by far the most effective cue and Jenkins (1961) who claimed that cues for stress perception in order of importance were pitch, timbre and loudness. Frost (2011) conducted a perceptual test of three correlates of stress – pitch, duration, intensity – to assess their individual importance and found F0 to be the most reliable, too.

2.1.5 Rhythm Metrics

The first efforts to quantify rhythm focused on variation in vowel and consonant duration, following Dauer's idea (Dauer 1983, 1987) that speech rhythm can be captured by measuring relative consonantal and vocalic variability. Ramus, et al. (1999) devised three measures (so-called *rhythm metrics*) that were supposed to provide "an implementation of the phonological account of rhythm perception":

- %V – the proportion of vocalic intervals within each sentence
- ΔV – the standard deviation of the duration of vocalic intervals within each sentence
- ΔC – the standard deviation of the duration of consonantal intervals within each sentence

After the metrics were calculated they were projected in three different Cartesian diagrams – (%V, ΔC), (%V, ΔV) and (ΔV, ΔC). The following studies (Grabe & Low, 2002; Low, Grabe, & Nolan, 2000) employed a new measure called Pairwise Variability Index (PVI). Two versions were calculated, raw rPVI and a normalised version nPVI which was devised to normalize for speech rate differences. The metrics were originally applied to Singapore English and then to a group of 18 languages. Due to a considerable overlap between the classes, the results support only a weak distinction between stress-timed and syllable-timed languages. Some of the rhythm metrics were considered unreliable because of their correlation with speech rate (e.g. Barry et al., 2003). Dellwo and Wagner (2003) normalised the rhythm metrics for rate by dividing the standard deviation of interval duration by the mean, using a measures termed VarcoV for vowels and VarcoC for consonants. White & Mattys (2007) showed that two measures – VarcoV and %V – were more more resistant than other measures, but not completely inert to speech rate variation than the others.

Rhythm metrics have been consequently used for measuring of L2 speech and the influence of L1 upon L2 (Low, Grabe, & Nolan, 2000; Gut, 2003; Lin and Wang, 2005; Carter, 2005; Whitworth, 2002). As there are too many aspects to consider, different rhythmic properties of L1 and L2 and the degree of non-native accent being the main ones, "studies of the influence of L1 on L2 production are intrinsically difficult to interpret" (White & Mattys, 2007). In the study of Carter (2005), the only informative measure for rhythmically distinct first and second languages was nPVI-V. The results for nPVI-V were intermediate for children with Spanish as their L2 (low nPVI-V) and English as their L2 (high nPVI-V). This was explained by much smaller extent of vowel reduction in the English of native Spanish speakers, who lack this feature in their L1.

White & Mattys (2008) observed several caveats challenging the applicability of rhythm metrics. The main objection is that rhythm metrics they were applied to read sentences only, not to spontaneous speech that is characterized by speech rate variations, hesitations and dysfluencies. The results of rhythm metrics for the same scripted versus unscripted sentences in White & Mattys (2008) proved that speaking styles are an obstacle that the measures are unable to deal with.

2.1.6 Rhythm metrics criticism

Rhythm metrics became subject to considerable criticism. According to Kohler (2009), speech rhythm research measured inappropriate properties that did not capture language rhythm. Moreover, the rhythm metrics completely disregarded the listener, which led Kohler to state that "speech rhythm is different from, and goes beyond, phonology-driven speech timing". Kohler (2009) suggested movement from instrumental to perceptual evaluation of speech rhythm:

> "Before physical measurement variables in speech production can be related to rhythmical patterns in a scientifically insightful way the type and degree of rhythmicity in the data needs to be evaluated perceptually by the competent language user. [...] It is only then that acoustic or articulatory and physiological measures can be seen as the physical exponents of rhythmic categories in speech interaction in different languages."

As argued by Kohler, there are four variables patterning rhythmicity in speech that need to be included in the new rhythm research paradigm for a comprehensive account of the speech

rhythm: fundamental frequency (F0), syllabic duration (rate), syllabic energy (loudness), and spectral dynamics. Kohler's new research paradigm, which considers the speaker, the listener and communicative function, "assumes no surface isochrony, measurable in speech production, but gives the listener the key role in deciding on what constitutes rhythmic regularity." In this way Kohler (2009) shifts focus to perceptual studies, drifting away from calculations of acoustic parameters to perceptual experiments which concentrate on the listener rather than the speaker.

2.1.7 English rhythm difficulties

It has been shown that non-target-like prosody is notoriously persistent even in advanced learners (e.g., Grosser, 1993), and it is often seen as one of the main stumbling blocks for L2 learners (e.g. Barry, 2007). Mastering the main linguistic principle of English rhythm, the contrast between stressed and unstressed parts of the discourse, seems to be one of the main challenges for L2 English speakers:

> Learners of English surely have problems not with stress, but the lack of it. Those who learn to speak clearly and stress everything have difficulty with unstressed syllables and words, especially in reconstructing function words from their fragmentary weak forms […] Relatively few foreign speakers of English—even if they otherwise appear to be highly proficient—are able to draw the correct inferences from the suppression of stress. (Knowles 1995: 288)

As noted by Chela-Flores (1994), the inability of EFL speakers to produce native-like English speech rhythm might also stem from excessive aim for meticulous pronunciation:

> English rhythm has been examined from a number of perspectives and we are now beginning to understand the true nature of the problems that it presents for EFL learners. We no longer attribute the difficulty to the theory of stress-timing and syllable-timing on which most teaching methodologies for pronunciation have been based […] the failure to make sufficient difference in length between the vowels in stressed and unstressed syllables seems to be the basic cause of difficulty with English rhythm among non-native speakers of English. […] A syllabic rhythm might be the result of too much emphasis on the pronunciation of each unit in an utterance and not producing adequately lengthened and shortened syllables in chunks. (Chela-Flores 1994: 235–6),

2.2 Speech rhythm perception

After failing to produce reliable results using rhythm metrics, numerous studies pointed in the direction of the speech perception in the processing of language rhythm (e.g. Kohler, 2009). If the speech rhythm is too elusive to yield to objective measuring tools, perhaps it is the time to turn to the listener whose perception plays a crucial role in the processing of speech rhythm. Subjective isochrony of speech rhythm was already observed, although not supported by any experiment, by Classe (1939): "A certain irregularity of syllabic distribution will disturb 'objective' isochronism, but not necessarily 'subjective' isochronism."

The listener actively participates in the speech processing, as noted by Handel (1989, p. 449): "The acoustic wave induces us to hear a rhythmic pattern, but the acoustic wave does not directly signal that pattern." These results motivate the future research to focus on the listener, so that a complex and insightful account of rhythm can be constructed.

As noted by Ramus (1999), the cues helping listeners perceive speech rhythm could emerge from "the succession of syllables, vowels, stresses, pitch excursions, energy bursts within a certain range of frequencies, or whatever occurs repeatedly in speech that the human ear can perceive."

2.2.1 Early language rhythm discrimination

There has been a considerable attention given to the acquisition of the language in early stages in life. A growing number of psychologists and phoneticians focus on early language discrimination to reveal the mechanisms involved in the process. Ability of newborns and infants to discriminate native from non-native speech has been demonstrated by Mehler et al. (1986; 1988), Bahrick and Pickens (1988), Jusczyk et al. (1993), Moon et al. (1993), Bosch and Sebastián-Gallés (1997) and Dehaene-Lambertz and Houston (1998).

The experiment of Jusczyk et al. (1993) worked with six and nine months old infants. It asked the question whether the infants will be more sensitive to their native language rhythmic pattern than to non-native ones. In English, majority of words are stressed on the initial syllable. The results showed preference of the infants to strong/weak stress patterns as opposed to the reverse weak/strong stress patterns. This was the case even when the speech input was low-pass filtered, which indicates that their preference is caused by the prosodic structure of the words. Jusczyk et al. (1993) concludes with the assumption that "attention to predominant stress patterns in the

native language may form an important part of the infant's process of developing a lexicon."

Besides the native/non-native language differentiation, it was demonstrated that newborns are able to discriminate between the languages belonging to different rhythm classes: between stress-timed and syllable-timed languages (Mehler et al., 1988; Moon, Cooper & Fifer, 1993) and between stress-timed and mora-timed languages (Nazzi et al., 1998; Ramus et al., 2000). By five months of age, infants can even distinguish the rhythms of their native language from other languages in the same rhythmic family (Nazzi, Jusczyk, & Johnson, 2000).

2.2.2 P-center

Central to the understanding of the perceptual nature of speech rhythm is the identification of the perceptual center which is the temporal reference point at which a syllable is perceived to occur (Cooper, Whalen, & Fowler, 1986; Fraisse, 1974; Hoequist, 1983; Marcus, 1981; Morton, Marcus, & Frankish, 1976).

The term 'P-center', or 'perceptual center', was coined by Morton et al. (1976) to refer to "the locus in a word that must be equidistant, temporally, from corresponding loci in surrounding words in order for the sequence to sound isochronous to a perceiver." The P-center is also explained by Morton et al. (1976) as "the psychological moment of occurrence of a word." They claimed that listeners evaluate the timing of the word sequences based on reference points occurring within each word, and their results indicate that the longer the acoustic duration of the initial consonant in the syllable, the longer the interval between the acoustic onset of the word and the location of its reference point, the P-center. However, Morton et al. (1976) were unable to discover any identifiable acoustic marker of the P-center.

Morton et al. (1976) also hypothesized that the P-center is independent of context, meaning the acoustic features and timing of temporally nearby sounds. According to this context independence hypothesis, the location of a sound's P-center is fixed and does not depend on the events that precede or succeed it. However, after several decades of research attempting to localize the P-center, its precise location and its corresponding acoustic representation are still the subject of debate.

One of the claims explaining the difficulty of finding isochrony in the speech signal was that stress timing is largely an imposition by a listener, not by a talker (for example, Coleman, 1974; Lehiste, 1973). This proposition is backed up by the study of Morton et al. (1976), in which the

participants failed to find isochrony in evenly timed sequences of spoken digits. Although the acoustic onsets were temporally equidistant, listeners needed to introduce systematic departures from acoustic isochrony in order to perceive the sequences as isochronous. These acoustic departures were precisely those that participants created in the study of Fowler (1979) when asked to produce an isochronous sequence. These findings suggest that listeners judge isochrony based on acoustic information about articulatory timing.

Referring to the previous studies of Rapp (1971) and Allen (1972), Morton et. al. (1976) point out that P-center can be thought of as a production center, too. Allen (1972) asked participants to tap their fingers or match an auditive pulse to given syllables of an utterance. The results showed that metronome beats preceded the vowel onset by an amount positively correlated with the length of the prevocalic consonant. By placing the stress beat very often within the acoustic realization of the stressed syllable's prevocalic acoustic signal, Allen (1972) challenged the assumption that P-center corresponds to the articulatory onset of a stressed syllable.

Fowler (1979) also suggests that P-centers are connected directly to gestural events in speech production, but the acoustic correlates of these events are not straightforwardly reflected in the acoustic signal due to the complexity of the articulatory to acoustic mapping. According to Fowler, this is the reason why the studies on speech isochrony and P-centers failed to locate isochronous events in the acoustic signal.

2.2.2.1 Rhythm adjustment method

Rhythm adjustment was first described in detail by Marcus (1981), and it is the most commonly used method for measuring P-centers (see, e.g., Cooper, Whalen, & Fowler, 1986; Harsin, 1997; Pompino-Marschall, 1989; Scott, 1998). According to this method, sequences are composed of cyclic repetition of just two sounds, the base sound and the target sound, i.e. base–target–base–target etc. While the base–base interval is fixed, the base–target interval is adjustable (see Figure 1). The repeating pattern is not perceptually isochronous at first. The participants are asked to adjust the timing of the test sound within the cycle until the point of subjective isochrony is reached and consecutive P-center-to-P-center intervals will become equal.

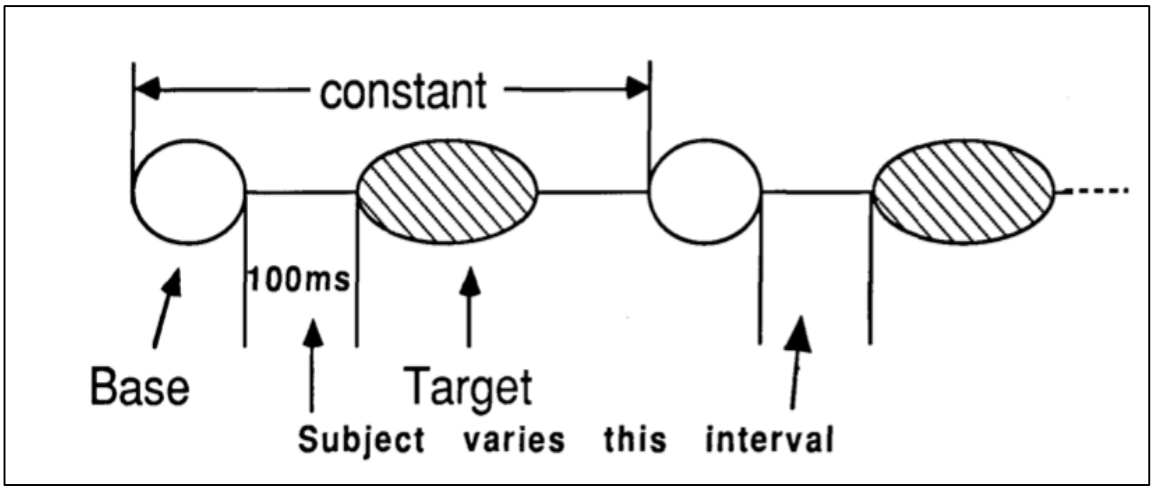


Figure 1. An illustration of the rhythm adjustment method as in de Jong (1992). Participants are presented with the sequence of alternating bases and targets. They adjust the relative timing of the target until the sequence becomes perceptually isochronous.

As pointed out by Barbosa et al. (2005), speech perception is linked with speech production in the case of speech synchronization tasks. Port (2003) suggest there is a single physiological mechanism behind the periodical production of acoustically prominent places in speech and production of the perception pulses in brain that are attracted by them. In research practice, speech production and speech perception are therefore treated independently as autonomous systems (Howell, 1988; Pompino-Marschall, 1989; Scott, 1993) or as closely interconnected mechanisms (Fowler, 1979, 1983; de Jong, 1992, 1994; Morton et al., 1976).

2.2.2.2 Speech perception approach

Research on the P-center is generally conducted from either the perspective of speech perception or speech production. In studies that used the speech perception approach (Cooper et al., 1986; de Jong, 1994; Marcus, 1981; Pompino-Marschall, 1989; Scott, 1998), participants listened to isochronous sequences of syllables with different initial consonants and rimes. The studies used either simultaneous or interpolating metronome beats as a temporal reference during the experiment, and the syllables were aligned in such a way that the inter-onset intervals were constant. Participants were then asked to make judgements whether the intervals between syllables were isochronous or not, either between subsequent syllables or between syllables and the metronome. They were then allowed to adjust the positions of the syllables until they could perceive the intervals as isochronous. These newly adjusted positions were then determined as the locations of P-centers in particular syllables.

2.2.2.3 Speech production approach

In studies that used the speech production perspective (e.g., Barbosa et al., 2005; Cummins & Port, 1998, 2009; de Jong, 2001; Hoequist, 1983; Tajima & Port, 2003) participants produced a sequence of syllables while synchronizing with a regularly occurring visual or auditory signal. The location within the produced syllables that coincided with the cue was identified as the P-center.

The difference of the pulse placement for syllables starting with voiced and voiceless consonant was found in speech production studies of Lindblom (1970) and Allen (1972). Their participants were asked to produce speech synchronously with equidistant metronome pulses. They found out that in syllables with voiceless consonants the pulse tends to occur 30-70 milliseconds earlier than with voiced consonants.

The study of Lidji et al. (2011) included a tapping task which is commonly used in research of timekeeping mechanisms. Participants were asked to tap along with the rhythm they perceived in English and French utterances. The results of the experiment suggest that long-term linguistic experience with a stress-timed language – i.e. English – can differentiate speakers' entrainment to rhythmic regularities. Speakers of a stress-timed language may also have greater expectations for rhythmic regularity in speech and may synchronize with the speech rhythm better.

Despite differences in experimental methodology and L1 of participants, most studies on P-center (e.g. Barbosa et al., 2005; Hoequist, 1983; Patel et al., 1999) assume that it is located close to the vowel onset within the transition between the syllable-initial consonant and the vowel. Nevertheless, objections to the speech production method have been raised by Villing et al. (2011):

> Nevertheless, due to the complex nature of the motor task involved in speech production, the variability between repeated productions of the same token, and the limitation to speech sounds only, methods using these tasks are not suitable for general P-center measurement.

2.2.2.4 Objections to P-center measurements

Sources of problematic issues regarding the research of P-centres were pointed out by Benadon (2013):

> The syntactical and acoustic complexities of speech may have precluded the formation of clearly defined rhythmic nuclei, leading different participants to assign onset markers to different sub-regions of the same speech sequence. […] The phonetic complexity of speech, coupled with shared knowledge of syntactical rules, could have led to more rigid interpretations of the speech rhythm.

The difficulty of assessing the rhythmicity of speech signal and finding exact location of the P-centre of a word may stem from the fact that contrary to metronome signal "the syllable beat is like a 'broad slur' rather than a single point in time" (Benadon, 2013).

Other less local parameters which might be responsible for the additional P-center drift include the shape of the decay ramp (Scott, 1998), the duration of the entire vowel or syllable (Fox & Lehiste, 1987; Scott, 1993 [experiment 8]), and the spectral envelope (Harsin, 1997; Howell, 1988; Pompino-Marschall, 1989).

Also, the majority of P-center research has been carried out with speakers of Germanic languages. In the studies, it was the vowel onset the seemed to be the main attractor for the perceptual center. However, as pointed out by Barbosa et al. (2005), this might be due to the fact that the syllable-initial position in Germanic languages commonly occupies a larger proportion of the syllabic duration and is subject to greater variability than in non-Germanic languages.

Nevertheless, there is general agreement that the location of the perceptual center is determined by the syllable's prevocalic portion, with longer consonantal onsets proportionally delaying the P-center (Cooper, Whalen & Fowler, 1986; Howell, 1988).

2.2.3 Resynthesis of speech

As mentioned in preceding sections, there has been growing evidence that language skills rely on rhythmic abilities that are not domain-specific (Gordon, Magne, & Large, 2011; Hausen, Torppa, Salmela, Vainio, & Sarkamo, 2013; Peter, McArthur, & Thompson, 2012) and can therefore be assessed using non-linguistic stimuli. Resynthesis of speech a procedure that converts speech to non-speech signal, retaining only some features of speech (e.g. rhythm, intonation curve) while eliminating others (phonotactics, lexical information). Through resynthesis, unnecessary linguistic features can be abstracted from speech in order to study its rhythmic and intonation properties in isolation.

First attempts to resynthesize speech for the purposes of language discrimination studies were carried out mostly in studies on infants. Atkinson (1968), Bonte (1975), and Ohala & Gilbert (1979) used pulse trains retaining the amplitude and frequency of the original speech signal. Bush (1967), Richardson (1973), Mehler et al. (1988), Nazzi et al. (1998), Bosch and Sebastián-Gallés (1997) and Dehaene-Lambertz and Houston (1998) used low-pass filtering to isolate prosodic cues from the speech signal. Low-pass filtering as a tool for eliminating segmental information and isolating prosody was criticised by Ramus et al. (1999), pointing out that "filtering does not allow one to know which properties of the signal are eliminated and which are preserved." Unfortunately, low-pass filtering "does not make any distinction between intonation and rhythm, and much information would be gained by separating these two components of the speech signal."

Therefore, an intonation hypothesis is formulated as a reaction to Nazzi et al. (1998), assuming that the language discrimination may have been based on intonation and not rhythm. Ramus et. al (1999) argued that while adults can rely upon lexical knowledge for language discrimination, infants do not have this opportunity and must therefore turn their attention on other cues that facilitate language discrimination. These can be differences in segmental repertoire, phonotactic constraints, or prosody. Importance of intonation for language discrimination shouldn't be neglected, as suggested by Maidment (1976; 1983), Ohala and Gilbert (1979), Willems (1982), and de Pijper (1983). Several studies demonstrated the ability to discriminate between some languages solely on the basis of their intonation: between English and Japanese (Ramus & Mehler, 1999), between English and French (Maidment, 1976, 1983) and between English and Dutch (Pijper, 1983). To ensure that the intonation cue is not used for language rhythm discrimination, Ramus (2002) proposes "to get rid of the intonation confound, that is to go beyond speech filtering and remove intonation from the stimuli."

Ramus et al. (1999) built their experiment on the supposition that "if one wants to test rhythm as a potential cue to discriminate between two languages, one should have stimuli that preserve as much as possible the organization of sequences of syllables and degrade as much as possible all alternative cues." In the experiment, English and Japanese sentences were resynthesized, delexicalizing the utterances and preserving only (1) broad phonotactics, rhythm, and intonation, (2) rhythm and intonation, (3) intonation and (4) rhythm. The first condition, termed *saltanaj* by the authors, consisted of replacing all fricatives with /s/, stop consonants with /t/, liquids with /l/, nasals with /n/, glides with /j/, and vowels with /a/. These phonemes are considered the most universal in their respective categories (Maddieson, 1984; Crystal, 1987).

The second condition, termed *sasasa*, "sasasa", consisted of replacing all consonants with /s/, and all vowels with /a/. The third condition, termed *aaaa*, consisted of replacing all phonemes with /a/, creating an intonation curve which sounded "like one long /a/, varying continuously in pitch." The fourth condition, termed *flat sasasa,* was similar to the first *sasasa* condition except that all sentences were synthesized with a constant fundamental frequency at 230 Hz. Their results showed that the listeners could do without any intonation (*aaaa* condition 3), and that syllabic rhythm was a sufficient cue for the speech rhythm discrimination (*flat sasasa* condition 4). Ramus et al. (1999), in which stress was signalled only by pitch excursions and duration, suggest the inclusion of amplitude in future speech rhythm research, assuming that it would "make it possible to analyse separately the respective roles of rhythm due to the succession of syllables and rhythm due to amplitude."

The study of Ramus et al. (2002) was designed to test the rhythm-class hypothesis by perceptual experiment comparing pairs of languages believed to have similar or different rhythmic properties. The properties of a language other than rhythm were eliminated by using only *flat sasasa* (condition 4) in the experiment. Through this resynthesis, all consonants were replaced by /s/ and all vowels were replaced by /a/. The fundamental frequency was ignored and replaced by a constant one at 230 Hz. Instead of giving out explicit instructions causing unnecessary bias in the listeners, the subjects were informed that they were to distinguish two exotic languages. They were presented with three sentences – two sentences of the same language as a context, and the third sentence either in the same language or in a different one. To make sure that the discrimination task was not carried out on the basis of the differences between the length of the utterances, the duration of the sentences was compensated for by multiplication. Their results confirmed that all languages belonging to different rhythm class (syllable-timed or stress-timed) were discriminated significantly above chance level. Ramus et al. (2002) conclude that "perceptual experiments investigating the discriminability of languages' rhythm by naive listeners should be the yardstick by which theories of speech rhythm will be measured."

2.3 Speech and music

2.3.1 Common origin theories

Speech and music are universal among human cultures. The relationship between them has long interested scholars across a broad range of disciplines, from linguistics to neuroscience. Both Jean-Jacques Rousseau and Charles Darwin supported the view that music and language share common origins. In his publication on the origin of language, Rousseau proposed the idea that the first languages were sung, not spoken (Rousseau, 1781/1993). Darwin elaborated upon this idea, suggesting the idea that music evolved from the love calls produced during the reproduction period to charm the person of the opposite sex: "musical notes and rhythm were first acquired by the male or female progenitors of mankind for the sake of charming the opposite sex" (Darwin, 1871/1981, p. 336).

Herbert Spencer, a prominent 19th century philosopher, anthropologist and sociologist, also favoured a common origin of music and language, and constructed a physiological theory to explain their common primary function to express emotions (Spencer, 1857). Various evolutionary theories suggest that music and speech could have had a common origin in an early communication system represented by holistic vocalizations and body gestures (Mithen, 2005), which is supported by the claim that music plays a crucial role in social interaction, especially between the mother and the infant (Threhub, 2003).

However, current research proposes differing views on the theories of common origin. Different views on their origin have proposed that either music might be a by-product of language (Pinker, 1997), language could have originated from music (e.g. Falk, 2004; Fitch, 2010) or language and music could have originated from a common cognitive domain (Brown, 2000).Whether their origins are identical or not, both music and speech can be defined as auditory communication systems that utilize similar acoustic cues for many purposes, for example for expressing emotions (Juslin and Laukka, 2003).

2.3.2 Functions of speech and music

Although both music and language are human universals present in all human societies, their main functions differ. Music facilitates expressing emotions and thereby ensures social bonding (Boucourechliev, 1993). In contrast, language is a tool for communication, storytelling and

planning. Over the course of human evolution, language lost the isomorphism between sound and meaning to become symbolic, containing signs which bear no readily recognizable physical resemblance to what they signify. This gradual process from the iconic to symbolic representation of language sign is seen in the almost complete disappearance of onomatopoeias from present-day languages.

2.3.3 Speech and music processing

Both music and speech are complex processing systems that work in close relationships with attention, memory, and motor abilities. Both of them comprise several levels of processing: morphology, phonology, semantics, syntax and pragmatics in language and rhythm, melody, and harmony in speech. Both of them are auditory signals that are sequential in nature, unfolding in time, governed by the rules of syntax and harmony. And sounds of both speech and music are built on the same acoustic parameters of frequency, duration, intensity, and timbre. The utilization of common mechanisms in language and speech has been suggested in syntactic processing (Patel, 1998, 2003a, 2008) as well as in melodic and rhythmic organization (Lerdahl and Jackendoff, 1983; Jackendoff, 1989). Both of them rely on the ability to integrate discrete acoustic events into a coherent perceptual stream according to specific syntactic rules (Patel, 2003).

2.3.4 Musical rhythm perception

The term *musical rhythm* refers to rhythms commonly used in Western music, such as those described in formal music theories (Cooper & Meyer, 1960; Lerdahl & Jackendoff, 1983; Yeston, 1976). These constitute a subset of all rhythms that exist all over the world and can be characterized by several typical rules. The rhythms of Western music are constructed around a regular pulse, with the majority of notes falling on beats rather than between beats. Note durations fall into several categories, with longer durations representing multiplications of shorter durations. Fraisse (1982) examined the frequency of occurrence of different note durations in a classical music repertoire and found out that the majority of pieces contained only two note durations in a ratio of 1:2. Although this ratio is considered to be the most dominant (Lerdahl & Jackendoff, 1983; Stoffer, 1985), other ratios such as 1:3 are commonly found in all kinds of Western tonal music. These note durations do not follow each other randomly but fit into structures governed by relative timing, in which each event related to all other beats in the sequence (Povel, 1981). Beat or pulse refers to a series of regularly recurring

psychological events that arise in response to the musical rhythm (Cooper and Meyer, 1960; Large, 2008; Grahn, 2012). It tends to correspond to one particular hierarchical level of the theoretical description of a rhythm (Figure 2).

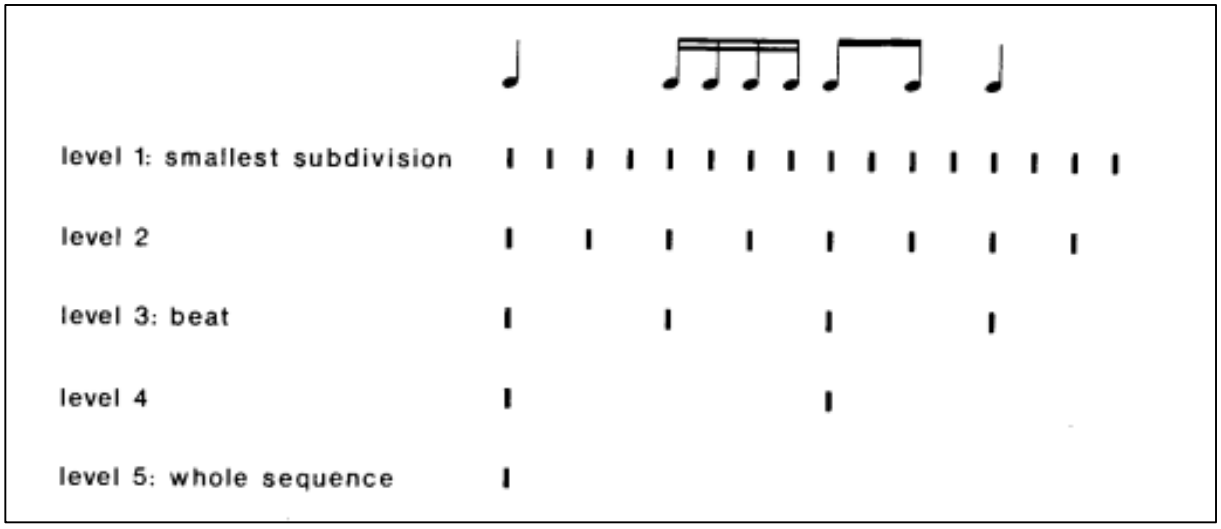


Figure 2. The musical rhythm is organized around a regular beat, each duration is twice as long as the one of shorter duration, and the note durations fit into a binary hierarchical structure (as in Drake, 1993)

It has been proved that perception of temporal regularities in the environment has a privileged psychological status over the perception of irregular sequences (Essens & Povel, 1985; Povel & Essens, 1985). Although musical rhythms may be subdivided in various different ways, researchers frequently consider binary subdivisions to be more perceptually salient than ternary or other subdivisions (Povel 1981; Jones, 1987, 1990).

2.3.5 Neurophysiology

Musical aspects of language such as rhythm, timbre and melodic contour are the central means of communication especially in infant-directed speech, and it has been shown that newborns show largely overlapping neural activity to infant-directed speech and to instrumental music (Kotilahti et al., 2010).

Compared to non-musicians, musicians show structural differences in the brain (Pantev et al. 1998; Gaser & Schlaug 2003), with larger grey matter volume in regions that are necessary for playing music. These areas include particularly motor, auditory and visuo-spatial regions (Gaser & Schlaug 2003) which are also indispensable for processing speech.

Numerous studies using modern functional neuroimaging techniques MRI, fMRI, or MEG have shown that the development of the perceptual, cognitive, and motor abilities, through years of intensive musical practice in the case of professional musicians, largely influences brain anatomy and brain function (e.g., Elbert et al., 1995; Schlaug et al., 1995a,b; Pantev et al., 1998; Schmithorst and Wilke, 2002; Gaser and Schlaug, 2003; Hutchinson et al., 2003; Bengtsson et al., 2005; Bermudez et al., 2009).

2.3.6 Musical and language training

Although speech production differs from music production in several dimensions (Hannon and Trainor, 2007), there have been theories suggesting that skills acquired through musical training may transfer to, and thereby improve, language related skills. However, the extent of the influence of musical training on speech processing and its underlying neurophysiological mechanisms remain a rich area to be explored.

Musical expertise acquired through formal music training have been associated with enhanced language skills (e.g., Besson, Schön, Moreno, Santos, & Magne, 2007; Milovanov & Tervaniemi, 2011; Schellenberg, 2005). The potential benefit of musical training for language skills may stem from the many shared anatomical and functional bases between the two domains (e.g. Patel, 2008). Researchers have pointed out the domain-generality of rhythm processing (Gordon, Magne, & Large, 2011; Hausen, Torppa, Salmela, Vainio, & Sarkamo, 2013; Peter, McArthur, & Thompson, 2012) which is indispensable for both music and speech perception.

Many acclaimed musicians have been training since a very young age, which resulted in a widespread belief that superior musical skills are associated with early training. Researchers have been trying to find out whether there is something specific about being exposed to this type of experience during an early, sensitive period of development. Thanks to years of training, musicians develop an expertise in their instrument or mastery over their voice. In the course of training, musicians gradually learn to attend to the fine-grained acoustics of musical sounds. These include pitch, timing and timbre, the three basic components into which any sound, whether music or speech, can be broken down. And studies of musical training confirm these functional advantages in rhythm tasks that result from enriched auditory experience. Musicians have been shown to be better at processing pitch, timing and timbre of music compared to non-musicians (Tzounopoulos & Kraus, 2009).

In Bailey & Penhune (2010) study, participants were categorized into early-trained and late-trained musicians depending on whether they started the musical training before or after the age of 7. The participants were matched on years of musical training, hours of current practice and experience. Early-trained musicians were better at reproducing the temporal structure of the rhythms in an auditory rhythm synchronization task. The findings of Bailey & Penhune (2010) thus support the idea of a sensitive period during the early years of childhood for developing sensorimotor synchronization abilities through musical training.

In the same way musical training enhances auditory perception, learning a second language might enhance the perception of rhythmic variation in language, such as sound duration and intensity, which are also present in the realm of music. Roncaglia-Denissen et al. (2016) found that studying a language with distinct rather than similar rhythmic features enhances their rhythm discrimination abilities.

Two principal areas of interest are prosodic and syntactic structure. Studies on prosodic structure examine the way duration, pitch and intensity are involved in structured rhythmic and melodic patterns in the two domains (e.g. Jusczyk & Krumhansl, 1993; Lerdahl & Jackendoff, 1983). Research on syntactic structure examines the way individual elements combine in hierarchical fashion to form coherent sequences in the domains of music and speech (e.g. Swain, 1997, Patel, 1998).

Musicologists and linguists have suggested, yet without satisfactory empirical evidence, that the prosody of a composer's native language can have impact on the structure of his or her instrumental music (e.g. Abraham, 1974; Wenk, 1987). A research on prosodic comparison of language and music even suggested there is "an empirical basis for the claim that spoken prosody leaves an imprint on the music of a culture" (Patel & Daniele, 2003).

Research provides ample evidence that musical and linguistic training results in improved perception of rhythm and that these two worlds appear to be closely intertwined. We will examine the role of musical training and practice in the experiment which will be dealt with in the following chapters.

2.4 Hypotheses and research questions

Before the experiment, following working hypotheses and research questions were formulated:

H1. Musically trained population will be more successful in detecting rhythm deviations than musically untrained population.

H2. Rhythm deviations in stressed syllables (or stressed beats) will be more evident than those in unstressed ones.

H3. Rhythm deviations directly before stressed syllables (or stressed beats) will be more evident than those before unstressed ones.

H4. Rhythm deviations in speech signal will be easier to identify than deviations in the purely rhythmical phrases where the linguistic information is missing.

H5. Rhythm deviations in both speech and non-speech phrases will be harder to recognize when they occur on later positions in the phrases.

R1. What is the shortest perceptible rhythm deviation in speech and musical phrases?

R2. Is the length of phrases a factor that influences the rhythm discrimination task?

R3. Which features of speech signal distract the listeners during the task of speech rhythm discrimination?

3. Method and Material

3.1 Recordings

In order to investigate perceptual differences between speech rhythm and musical – i.e. non-speech – rhythm, it was necessary to find appropriate speech stimuli and their non-speech counterparts.

3.1.1 Speech stimuli

For the speech stimuli, our goal was to choose various short phrases produced at a most natural speech rate by various native speaker. Recordings of five native English speakers – three female and two male speakers – were selected from the Prague Phonetic Corpus (Skarnitzl, 2010). For each speaker, three short English phrases were chosen. These phrases had the length of four, five and six syllables. All recordings were normalized for 32000 Hz sampling rate and 85% loudness. Also, three different short phrases – testELSA4, testJCA5 and testSF6 – with the length of four, five and six syllables were prepared for introductory training and instructions, none of which was used later in the experiment.

material		
file	text	syllables
ELSA4	*prove such a plan*	4
ELSA5	*short and medium range*	5
ELSA6	*the leaders of Russia*	6
IS4	*come amongst us*	4
IS5	*what was going on*	5
IS6	*certainly succeeded*	6
JCA4	*drinking sessions*	4
JCA5	*poverty wages*	5
JCA6	*against the chief justice*	6
SF4	*between his teeth*	4
SF5	*most unluckily*	5
SF6	*pulled an old pillow case*	6
SMA4	*about the case*	4
SMA5	*molesting children*	5
SMA6	*part in a rescue plan*	6
testELSA4	*peace-keeping troops*	4
testJCA5	*government spokesman*	5
testSF6	*family of wizards*	6

Table 1. Fifteen English phrases by five different native speakers (ELSA, IS, JCA, SF, SMA) were selected for the experiment. Three phrases with the length of four, five and six syllables were chosen for each speaker. Three additional phrases were used for introductory training.

3.1.2 Non-speech counterparts

These were created using a sample of two non-speech percussive sounds of woodblock. Two woodblock samples, high-pitched and low-pitched one, were used for the substitution of stressed and unstressed syllables, respectively. The speech stimuli were converted into two stereo channels. In one channel the original stimulus was preserved, and in the other the percussive sounds were aligned with the stressed and unstressed syllables of the phrases. The length of the percussive sounds was 25 milliseconds. It must be noted that the alignment was guided by the shape of amplitude envelope; the percussive sounds were aligned with the point of a rapid amplitude increase in the center frequency region, generally at or very near the vowel onset, which combined empirical evidence identifies as the syllable's perceptual center (Cummins & Port, 1998; Fowler, 1983; Scott, 1993; Scott, 1998).

3.1.3 Temporal modifications

For the speech stimuli, temporal modifications were performed using PSOLA algorithm in Adobe Audition software. PSOLA is a digital signal processing technique invented in 1980s which divides the speech waveform in small overlapping segments. In order to alter the duration of the signal, the segments are replicated or repeated multiple times, and combined through the overlap-add method.

One of the syllables within the speech phrase was chosen for the temporal manipulation, according to where the unnatural speech distortion resulting from the modification was the least audible. The variability of placements within speech phrases was observed, so that each possible placement was included in the set of fifteen phrases at least once (from 1st to 5th syllable). The original syllables that were subject to the modification varied in length from 68 to 189 milliseconds. To ensure that the modifications were not entirely inaudible or the identification became too simple, the syllables were lengthened by 150%. As the modifications were performed in the stereo channel including speech in one and non-speech stimuli in the other mono channel, the same method of modification were applied to both types.

TEMPORAL MANIPULATIONS				
file	location	original ms	result ms	diff ms
ELSA4m	3	162	243	81
ELSA5m	3	189	283	94
ELSA6m	1	98	147	49
IS4m	1	68	102	34
IS5mm	4	101	152	51
IS6m	5	100	150	50
JCA4m	1	162	243	81
JCA5m	2	121	182	61
JCA6m	3	130	195	65
SF4m	1	139	208	69
SF5m	3	139	209	69
SF6m	4	77	116	39
SMA4m	3	121	182	61
SMA5m	1	80	120	40
SMA6m	2	94	142	48
testELSA4m	2	122	183	61
testJCA5m	2	108	162	54
testSF6m	4	138	207	69

Table 2. Location of the temporal manipulations within the phrase (i.e. in which syllable the manipulation occurs), duration of the original and resultant item, and the difference in duration between the unmodified and modified stimulus (in milliseconds).

The fifteen recordings were put into pairs, separated by 1000 milliseconds of silence. Pairs of two kinds were created: they consisted either of two unmodified originals (value Same) or of the first unmodified phrase followed by its modified counterpart (value Different).

The stimuli were organised into three different blocks – speech, non-speech and cross-modal block. The speech block contained only pairs of speech items, and the non-speech block contained only pairs of non-speech items. Pairs in the cross-modal block consisted of speech items followed by their non-speech counterparts. Each block contained thirty stimuli. The first twenty stimuli in each block were unique items. As there were fifteen different items used in the experiment, the remaining five of the opposite same/different value were chosen to add up to twenty. The order of items in the experiment was designed so that the number of syllables in the phrases would alternate so as not to advantage or disadvantage any phase in the block. Also, the recordings of each individual speaker were spaced out in maximum distances, each occurring on every sixth position in the experiment. Same/different values were randomly assigned to the recordings in the block to ensure any potential for regularity or predictability

was eliminated. Remaining ten were control items used later for the calculations of intra-respondent consistency as an indicator of cognitive load. They were repetitions of ten previously tested items in the block, in the maximum distance in the experiment from the original. The same process was applied to non-speech and cross-modal block, with slightly altered order of control items and a recording of different speaker starting the block. Three tables in Appendix A show the complete order of items in individual blocks.

All pairs of phrases were introduced by a short sound signalling the start of a stimulus. The sound was located 500 milliseconds before the first stimulus separated by silence. Approximately 2.5-seconds-long tone that passed from low through high frequencies followed 1500 milliseconds after the second stimulus. The sound signalled the end of the first testing item, provided time for the respondents to mark their answer and prepared their ears for the next pair of stimuli.

3.2 Respondents

Eighteen healthy respondents (8 men, 10 women, mean age = 25.6 years, age range = 21–32 years) who had no history of neurological or psychiatric disorders were asked to participate in the perceptual experiment. To explore the influence of preceding musical training and current musical activities on the ability to hear rhythmical deviations, respondents were chosen so as to represent two groups of population. Approximately half of the respondents had no or very little musical training and were not currently involved in any musical activities, and another half of the respondents were active musicians with extensive musical training. English L2 skills of the respondents were also taken into account for further considerations. Keen participation of the subjects in the experiment was supported not by financial reward but by the respondents' genuine interest in testing their 'hearing abilities' and 'musical aptitude'.

3.3 Experiment

The recordings were played from a laptop, using noise-isolating (35 dB – 42 dB) earphones Etymotic Research HF5. The subjects were tested in quiet rooms without visual or acoustic distractions that would interfere with the concentration of the respondent.

3.3.1 Introductory training

The introductory training took place before every block and was designed to serve as a short

familiarization process for the respondents with the nature of temporal modifications. For each block they were handed one sheet of paper with a chart to mark down their answers. The sheet, included in Appendix F, contains three separate columns, one for each block. First, smaller chart serves for four stimuli intended for the introductory training, followed by a larger chart for thirty testing items. The items are numbered and two options ‘same’ and ‘different’ are represented by pictograms.

The short introductory training consisted of four training stimuli. The first training item played was without manipulation so that the respondents could get familiar with the pace of the experiment and the sounds surrounding the stimuli – introductory signal and closing sound. The second item contained temporal modification and the respondents were asked to concentrate on the difference between the two items. If required, the stimulus was played again. The modification from the second training stimulus was deleted from the third so that the respondents could notice the change that occurred. The respondents were then asked to find the temporal modification imposed on the fourth stimulus. The subsequent experiment was performed only after they became comfortable with the task of differentiation between modified and non-modified stimuli.

The experiment contained all thirty stimuli in a row and the respondents were told that it would be played as a whole, five-minute-long recording, without the option to repeat or pause it. Half of the respondents started the experiment listening to the speech stimuli, and the other half with the non-speech stimuli. The other – speech/non-speech – block followed second, and the cross-modal block came last.

Small talk was undertaken with the respondents between the three blocks to ensure that the cognitive load was not unbearable and they can have a little break from the intensive concentration. The reactions to the experiment were mostly centred on low self-confidence in the discrimination task. Majority of respondents spontaneously admitted insufficient capability to identify the temporal modification, although their results proved to be different. Most respondents also concluded that the speech block was much harder that the non-speech block, as speech signal presented them with many more things to divert their attention from mere rhythm to other domains such as melody and meaning. The cross-modal block was unanimously identified as the most difficult one.

SPEECH		NON-SPEECH		CROSS-MODAL	
file	same/diff	file	same/diff	file	same/diff
testSF6	S	testJCA5	S	testELSA4	S
testELSA4	D	testSF6	D	testJCA5	D
testELSA4	S	testSF6	S	testJCA5	S
testJCA5	D	testELSA4	D	testSF6	D

Table 3. Four testing items played before each block as part of the introductory training.

4. Results and discussion

In this chapter, the data collected from the perceptual test are analysed. In following sections, the influence of various variables on the results will be presented. In section 4.1, influence of the first block will be analysed. Section 4.2 will deal individually with unmodified and modified stimuli and their effect on the results. Section 4.3 will present results according to the scale of manipulation. Section 4.4 will analyse the results according to the presence or absence of stress in manipulated elements or in the elements directly following them. In Section 4.5, results will be linked to the length of the phrase (number of syllables in the phrase) and in Section 4.6 to the position of the temporal manipulations within the phrase. Section 4.7 will be concerned with the consistency of answers among listeners and the consistency of results for the first and second occurrences of individual items. Section 4.8 will present an analysis of the individual items reappearing in the blocks, and Section 4.9 will link the differences in results with the personal characteristics of the participants.

Mean success rate for all respondents in all blocks was 67.16%. The respondents achieved approximately the same results of 72.78% and 71.85% for speech block and non-speech block. As Figure 3 suggests, the cross-modal block presented a significantly more difficult task ($p = 0.006$), with only 56.85% mean success rate. The success rate across all individual respondents in all blocks is shown in Appendix B.

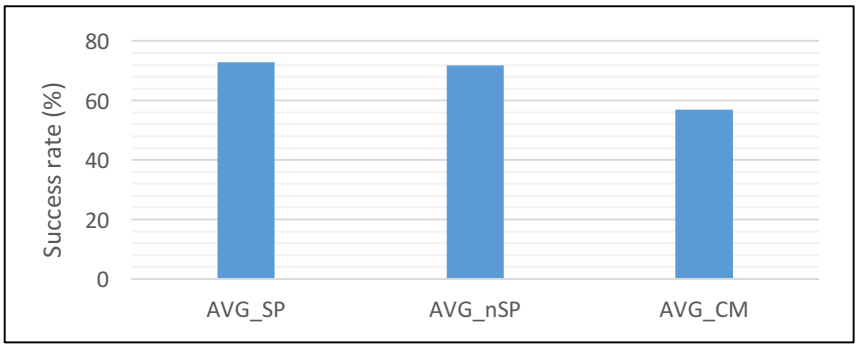


Figure 3. Mean success rates achieved in individual blocks – speech block (AVG_SP), non-speech block (AVG_nSP), and cross-modal block (AVG_CM)

4.1 Ordering: speech / non-speech block

To examine the influence of the type of stimuli in the first block of the experiment to the identification of rhythm deviations, one half of the respondents started the experiment with the speech block and the other half with the non-speech block. The results in Figure 4 illustrate the differences. Mean success rates for speech and non-speech were higher for the respondents starting with the speech block. The respondents were slightly more successful in cross-modal block when they started the experiment with the non-speech block, however, this difference was insignificant ($p = 0.6$).

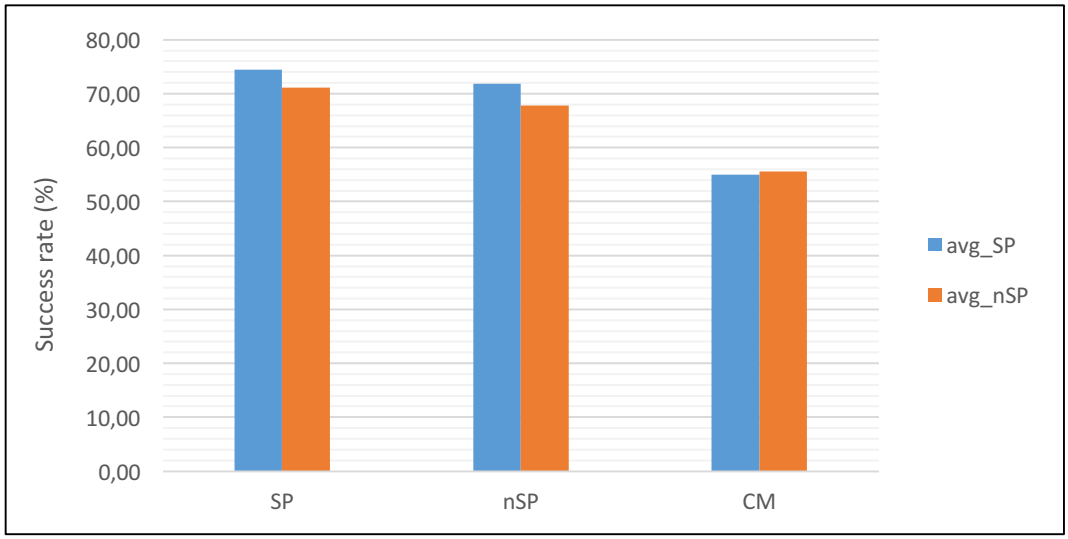


Figure 4. Mean success rate in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) of the participants who were first presented with the speech-block (avg_SP) and those who were first presented with the non-speech block (avg_nSP).

4.2 Value: Same / Different

Success rates for manipulated items was substantially lower than the success rate for the unmanipulated items in all blocks, except for the cross-modal block in which the success rate differed by only 1.39 %. The results are presented in Figure 5.

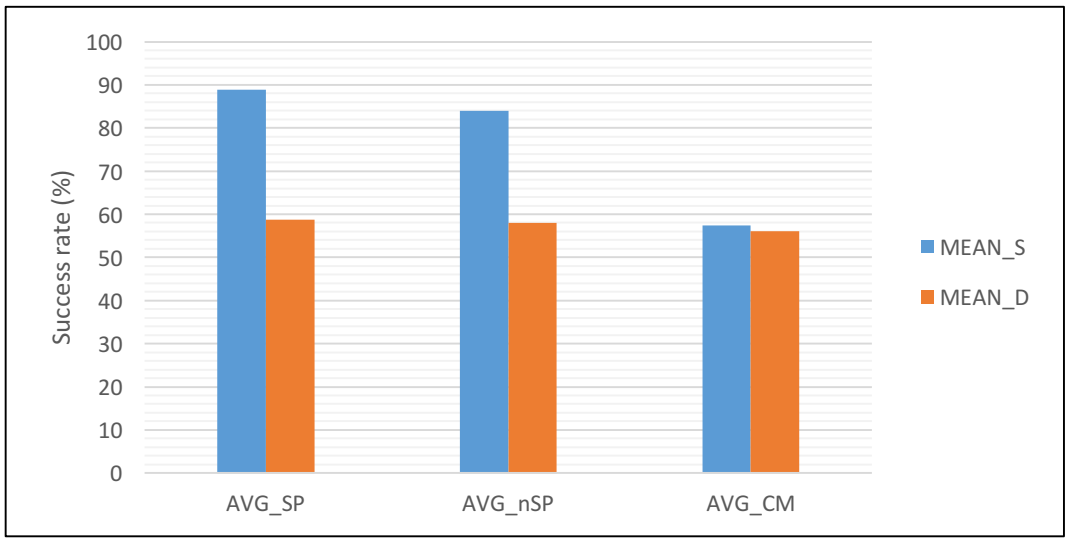


Figure 5. Mean success rates for temporally modified (MEAN_D) and unmodified (MEAN_S) items in speech block (AVG_SP), non-speech block (AVG_nSP) and cross-modal block (AVG_CM).

4.3 Scale of manipulation

The duration of manipulated intervals was variable, depending on the length of original syllables being modified ranging from 68 to 189 milliseconds. All the intervals were extended by 50%, and therefore the scale of manipulation of the items used for in the experiment ranged from 34 to 94 milliseconds. To evaluate the success rate for individual items in the experiment, the recurrence of ten controlling items in each block had to be taken into account. In all the sections that will deal with individual stimuli, the average of success rates of the two occurrences of a given stimuli in the block will be used.

4.3.1 Speech block

The calculations for stimuli in the speech block are presented in Figure 6. The mean success rate remains consistently above 58.68% – the overall average for manipulated speech items – when the difference between the original and the manipulated stimulus exceeds 65 milliseconds. The mean success rate for the stimuli in the intermediate scale of manipulation between 50-61 milliseconds remains below 50%.

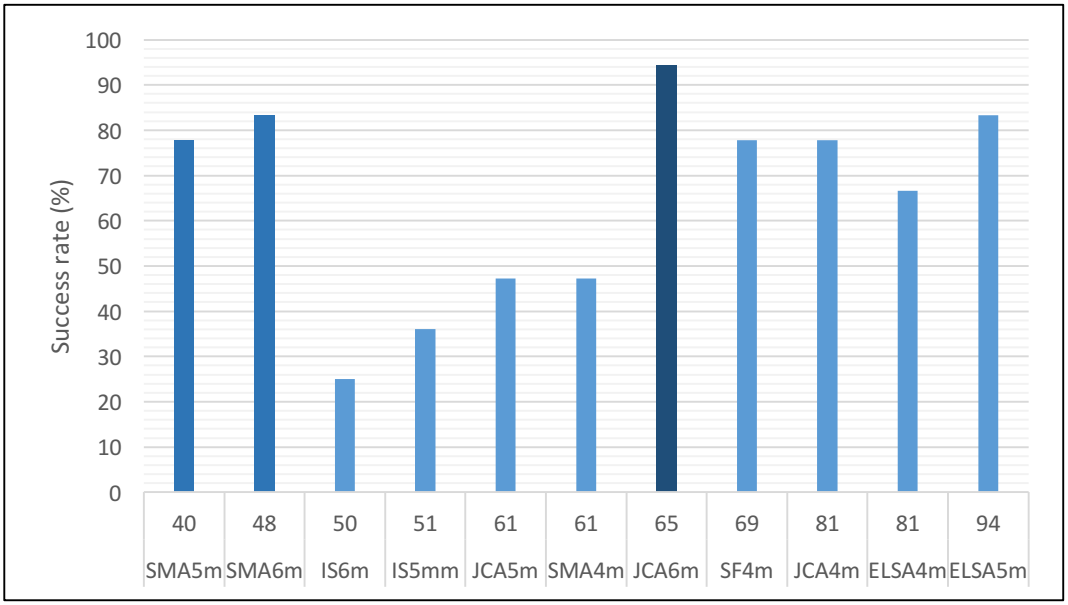


Figure 6. Mean success rates for temporally manipulated stimuli in the speech block with their scales of modification (in milliseconds) in an ascending order.

Several items defy expected progress that would regard the items with shorter manipulations less easily identifiable. The recording that was the easiest to identify in the speech block was not the one with the longest duration of manipulation. JCA6m was manipulated by 65 milliseconds and had the highest success rate out all the items in the speech block of 94.44%. Two items**,** SMA5m and SMA6m, with the shortest manipulations of only 42 and 48 milliseconds exhibit a relatively high success rate of 77.78% and 83.33% which are comparable to the items with the length of manipulation almost twice as long.

It was relatively difficult to identify the rhythm deviation in ELSA4m, the phrase with the second longest duration in the speech block with manipulation of 81 milliseconds, but only 66.67% success rate. The phrase *prove such a plan* is produced with a breathy articulation and heavy aspiration, which might have blurred the temporal exactness of the phrase and distracted the listeners from correctly identifying the manipulation. ELSA4m achieves markedly higher success rate of 83.33% in the non-speech form. This fact suggests that the speech signal might have clouded the rhythm of the phrase, the fact which was absent in the percussive non-speech phrase.

4.3.2 Non-speech block

The calculations for stimuli in the non-speech block are presented in Figure 7. For manipulated non-speech stimuli, the mean success rate remains consistently above 57.94% – the overall

average for manipulated non-speech items – when the difference between the original and the manipulated stimulus exceeds 61 milliseconds. The only exception, JCA6m with the manipulation of 65 milliseconds, was the most difficult for identification from the whole non-speech block with only 30.56% success rate

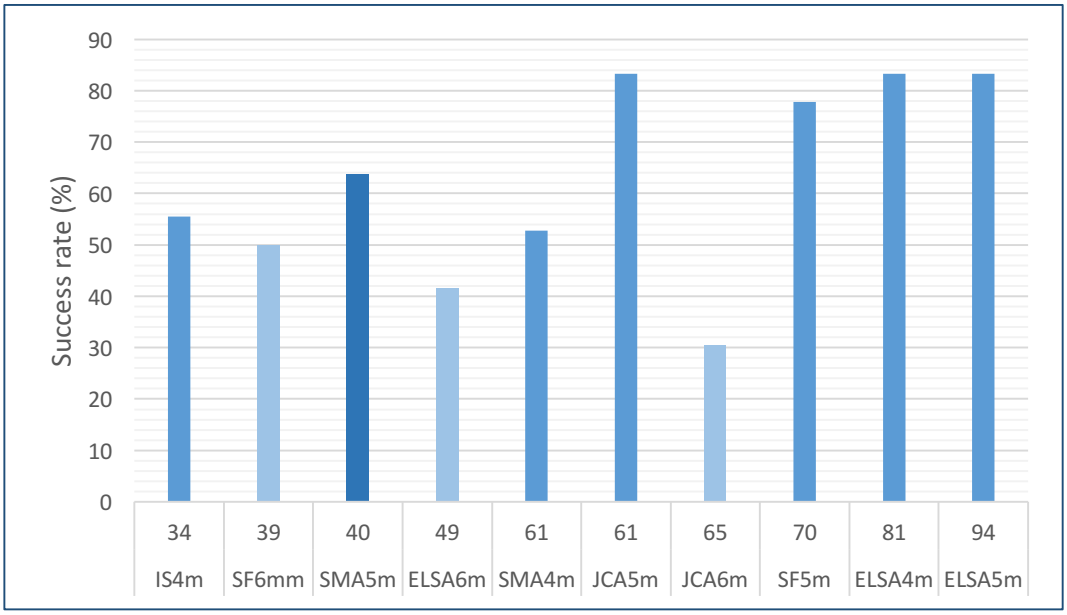


Figure 7. Mean success rates for temporally manipulated stimuli in the non-speech block with their scales of modification (in milliseconds) in an ascending order.

The three items with the lowest success rate in the non-speech block – JCA6m with 30.56%, ELSA6m with 41.67% and SF6m with 50% – have the length of six beats. Moreover, the higher the scales of modification of these six-syllable long stimuli are, the lower the success rate becomes.

4.3.3 Cross-modal block

Mean success rates for stimuli in the cross-modal block ranged between 38.94-66.67%. The development in Figure 8 is almost regular, with the mean success rate remaining consistently above 56.02% – the overall average for manipulated cross-modal items – with the scale of modification of at least 69 milliseconds.

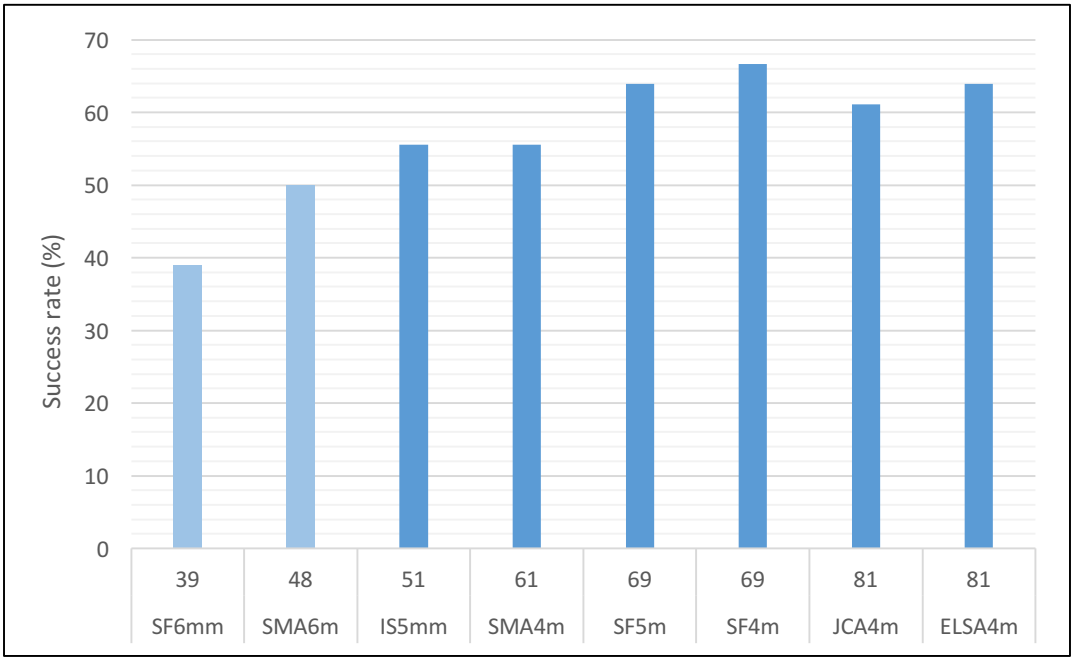


Figure 8. Mean success rates for temporally manipulated stimuli in the cross-modal block with their scales of modification (in milliseconds) in an ascending order.

Considering all three blocks at once, several tendencies can be observed. Higher-than-average success rates are reached at different scales of modification for each block. For the speech block, the success rate dramatically improves when the difference between the unmodified and modified item exceeds 65 milliseconds. For the non-speech block, the boundary is 61 milliseconds. And for the cross-modal block, the length of manipulation needs to reach at least 69 milliseconds. All three values fall into the interval of 60 to 70 milliseconds.

Complete comprehensive chart in Appendix C provides overview of mean success rates of all stimuli in all three blocks with the growing scale of modification.

4.4 Stress

The manipulations were performed on syllables that were either stressed or unstressed. In the non-speech domain, a higher and lower pitched metronome sound was used to represent stressed and unstressed syllable. Mean success rates for unstressed (65.86%) and stressed (65.59%) positions across all blocks were almost identical. The comparison of mean success rates for unstressed and stressed positions in individual blocks is shown in Figure 9.

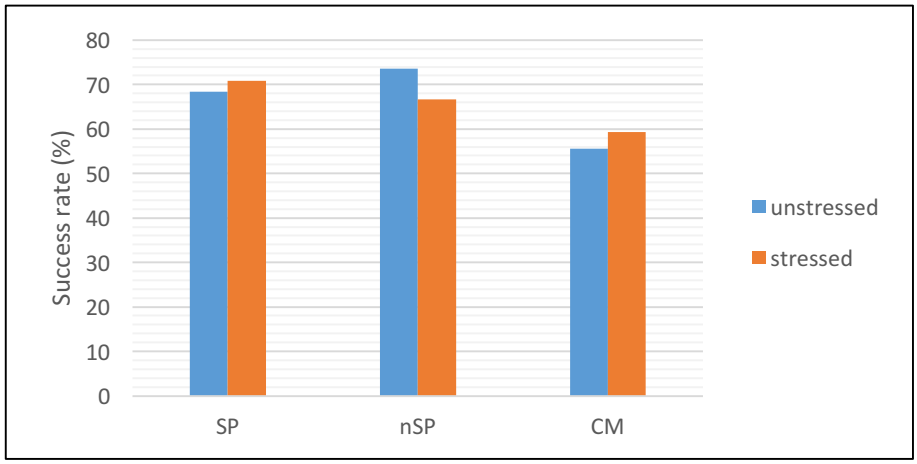


Figure 9. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) for the stressed and unstressed syllables/beats in temporally manipulated positions.

For the speech and cross-modal block, the success rate is higher for stressed positions, although the opposite is evidenced for the non-speech block. The identification of temporal manipulations was more successful for unstressed beats (73.61%) rather than stressed ones (66.67%).

Besides the target element that was subject to the modifications, the following syllable/beat was analysed too. Mean success rate reached 65.72% in positions when a stressed syllable/beat followed the target, but only 62.13% for an unstressed syllable/beat. The comparison of mean success rates for unstressed and stressed syllables/beats following the manipulated element in individual blocks is shown in Figure 10. The results are similar to the calculations for the target syllables in the preceding section.

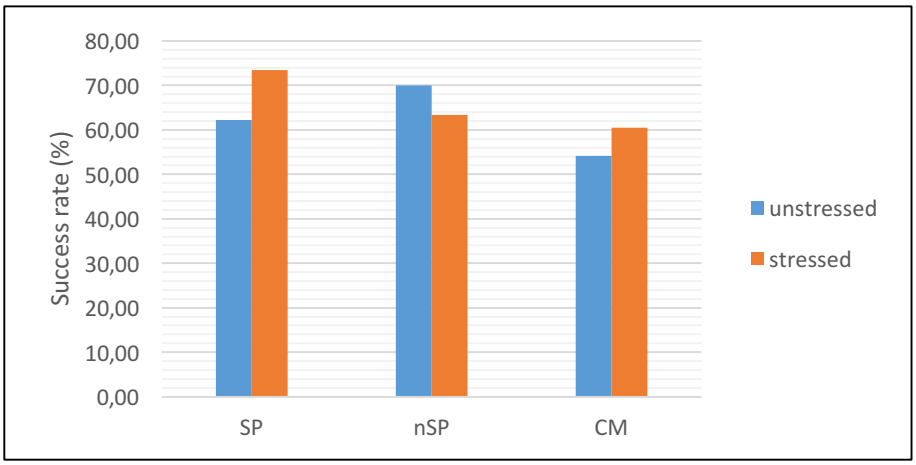


Figure 10. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) for temporally manipulated items according to the nature (stressed/unstressed) of the syllable/beat directly following the modified interval.

For the speech and cross-modal block, the success rate is higher for stressed positions. On the other hand, in the non-speech block subjects achieved better results in the identification task of unstressed beats.

4.5 Number of syllables in the phrase

In the experiment, three different phrase lengths of four, five and six syllables/beats were used. The difference between these three lengths in each block is illustrated by Figure 11. The mean success rates for stimuli with four or five syllables were slightly different for all blocks, with four syllable phrases more successful in speech (by 2.78%) and cross-modal (by 1.11%) block, and five syllable phrases more successful in the non-speech block (by 4.63%). In all three blocks, the mean success rates for the stimuli with the length of six syllables were significantly lower than the other two ($p = 0.002$).

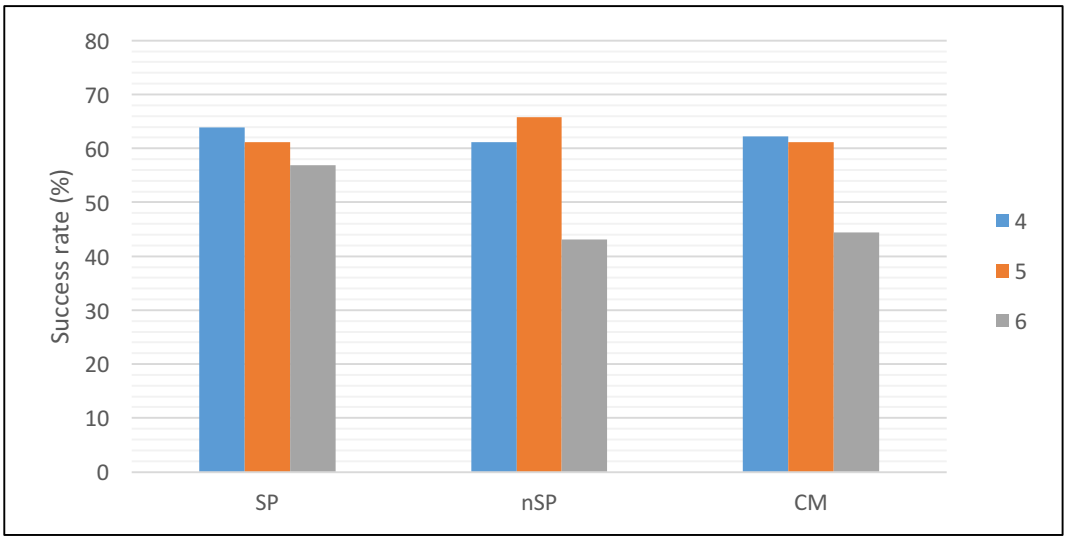


Figure 11. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) for temporally manipulated items grouped by the number of syllables/beats (four to six) in the phrases.

4.6 Position of the modification within the phrase

The maximum length of the phrase used in the experiment was six syllables/beats. There were five potential positions on which the temporal modification could occur, from the first to the fifth/syllable interval. As the Figure 12 shows, the success rate tends to decrease with later positioning of the modification in the phrase. The highest success rate of 83.33% was found for the second interval in non-speech phrases, and for the speech and cross-modal stimuli, the

highest success rates were in the first position (77.78% and 63.89%). In the speech and cross-modal block, the third position (67.29% and 62.22%) exhibited higher success rate than in the second position (59.26% and 50%). There is a steady decline in mean success rates for fourth and fifth position which was documented only for speech stimuli.

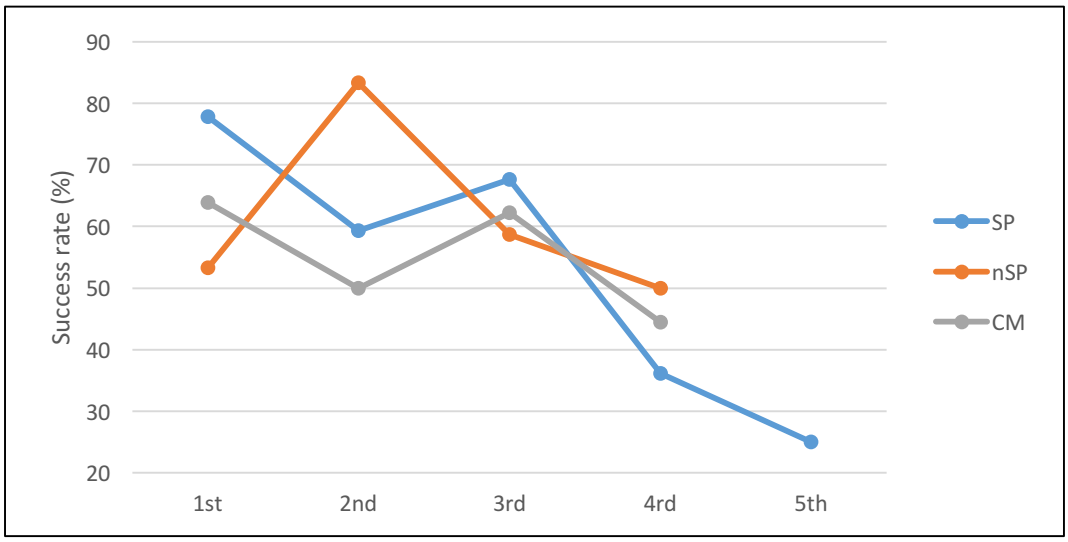


Figure 12. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) for temporally manipulated items according to the position of the modification (1st to 5th interval) within the phrase.

4.7 Intra-respondent consistency

The last ten items in each of the three blocks in the experiment were designed to provide information on intra-respondent consistency. Every subject displayed different degree of reliability on whether he or she can evaluate particular stimulus in the same way as before. The introduction of repeated stimuli at the end of each block served also as an indicator of cognitive load, suggesting whether the level of attention required from the respondents was not too high. As Figure 13 shows, intra-respondent variability did not exceed 50% for any of the respondents, with only two respondents (R8 and R14) with intra-respondent variability over 40%. The average intra-respondent consistency was 69.81% and the highest was 90% (R1).

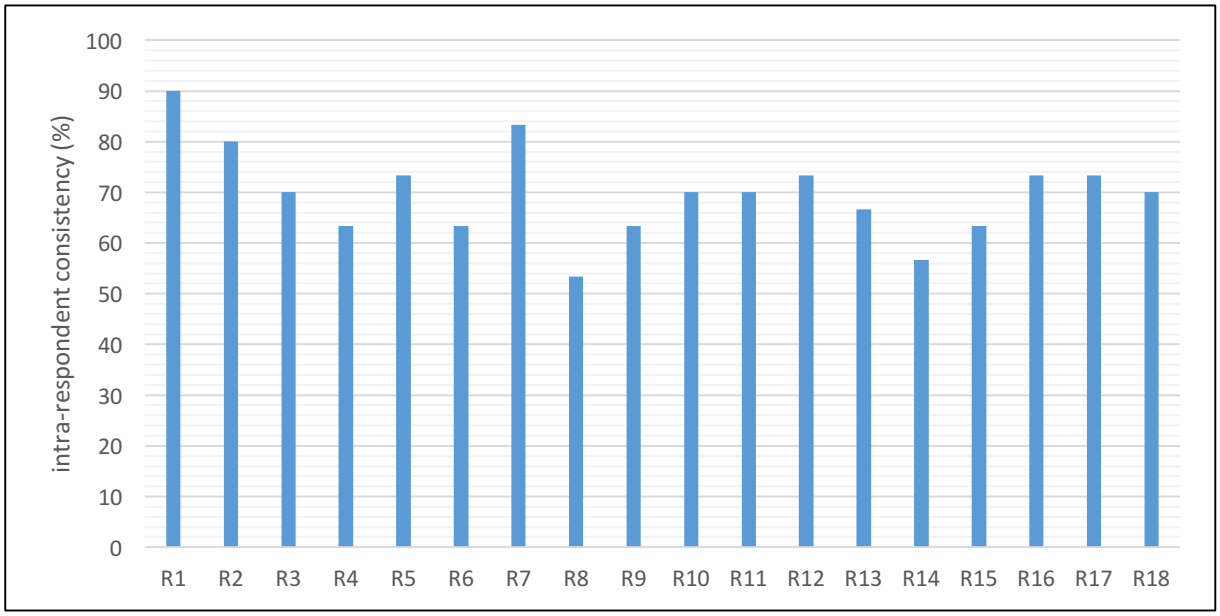


Figure 13. Intra-respondent consistency for individual respondents (R1 to R18) in the experiment.

Mean intra-respondent consistency was 73.33% for the speech block, 71.11% for the non-speech block and 65% for the cross-modal block. A complete chart with intra-respondent consistency for all speakers in all blocks is included in Appendix D. The lowest intra-respondent consistency was only 40% only in three cases, two out of which being in cross-modal block. For R1 and R12, the variability became non-existent in the speech block with the consistency of 100%.

4.7.1 First and second occurrences

The experiment contained ten items that occurred twice within each block. Mean improvement rate was calculated for the differences between the success rates for the first occurrence of a stimulus in the block and their second occurrence as controlling items at the end of each block. Items in the speech block displayed the worst deterioration of the second occurrences of average rate -10%, followed by the non-speech block with -6.67%, and the cross-modal block with the least deterioration of -3.89%. As seen on Figure 14, 15, and 16, only for nine stimuli out of thirty – two in the speech block, four in the non-speech block and three in the cross-modal block – did the success rate improve, the most dramatic being 44.44% of the non-speech, unmodified item SF4.

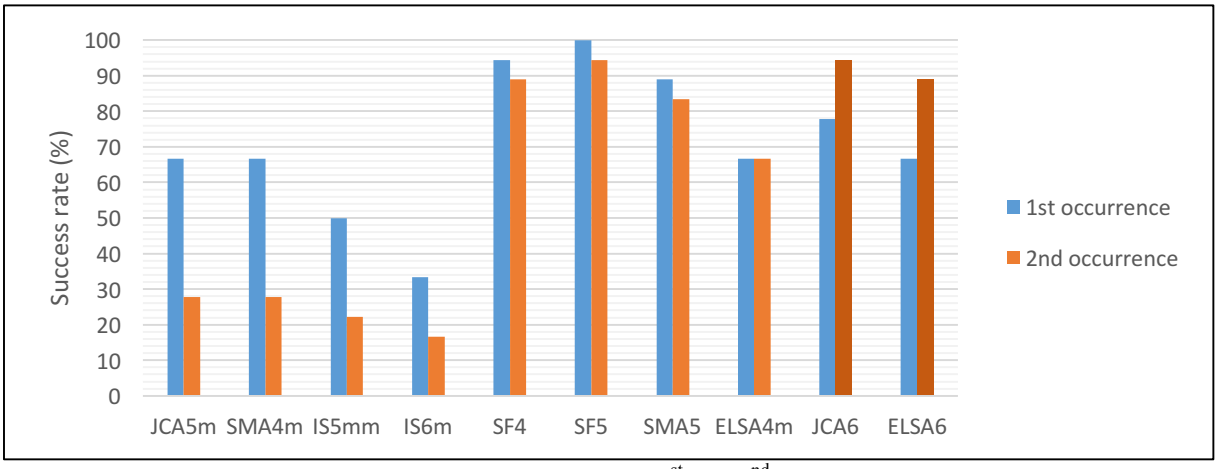


Figure 14. Differences in mean success rates between the 1st and 2nd occurrence of a stimulus in the speech block.

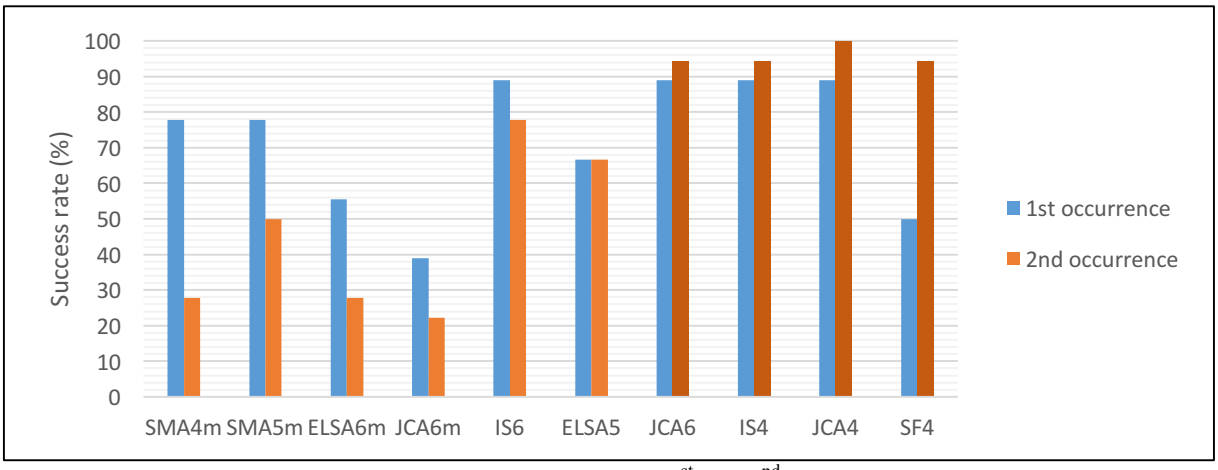


Figure 15. Differences in mean success rates between the 1st and 2nd occurrence of a stimulus in the non-speech block.

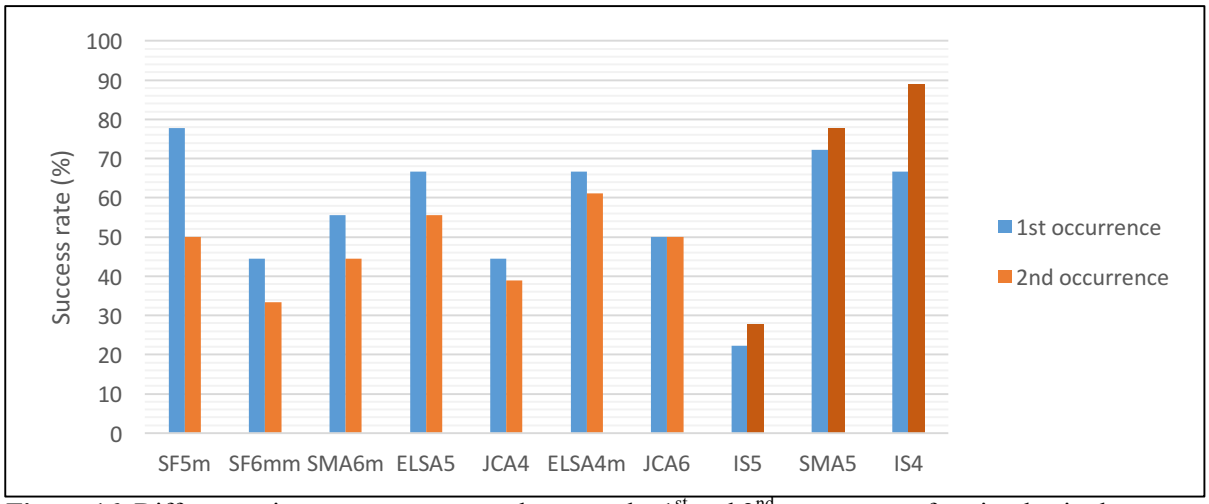


Figure 16. Differences in mean success rates between the 1st and 2nd occurrence of a stimulus in the cross-modal block.

4.8 Individual stimuli across blocks

Comparison of individual stimuli across various blocks was possible for several items that occurred more than in one block.

There were six stimuli that occurred both in the speech and non-speech block, as plotted in Figure 17. SMA5m, JCA5m and JCA6m achieved better success rates for the speech variant, with JCA6m possessing a considerable difference of 63.88% between the speech and non-speech block. JCA5m, SMA4m and ELSA4m displayed higher success rates for non-speech version, and ELSA5m, the phrase with the longest temporal modification of 94 milliseconds, had equal success rate of 83.33% in either block.

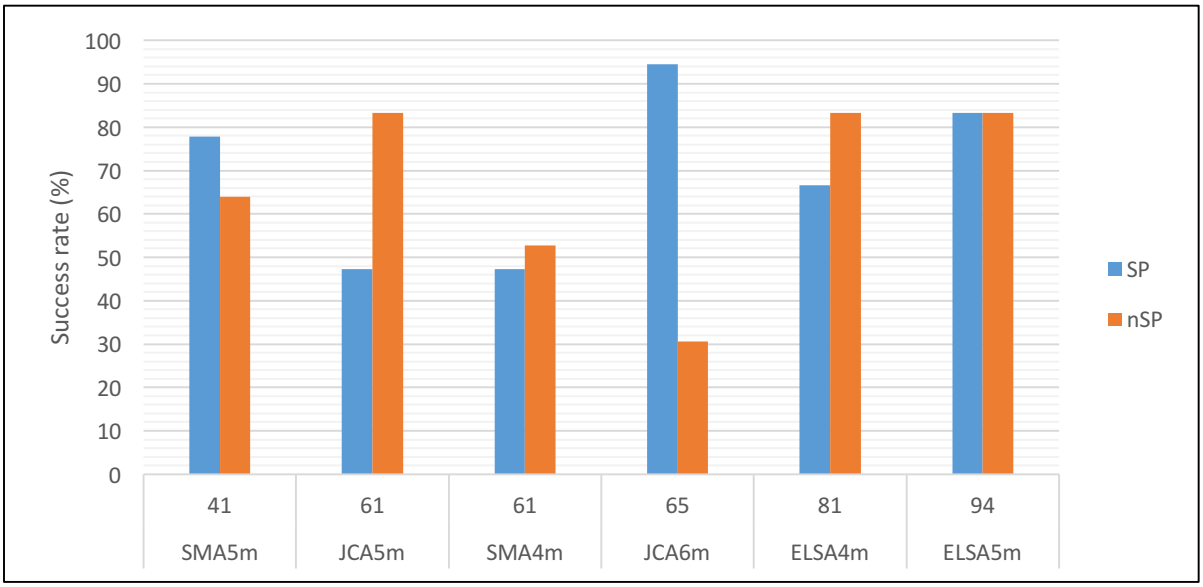


Figure 17. Mean success rates for temporally manipulated stimuli occurring both in speech block (SP) and non-speech block (nSP).

The differences in success rates for stimuli counterparts in speech and cross-modal block were more straightforward, as seen in Figure 18. For five out of six items, the success rate among cross-modal stimuli was considerably lower. Only for cross-modal instance of IS5m is the success rate higher than for its speech version.

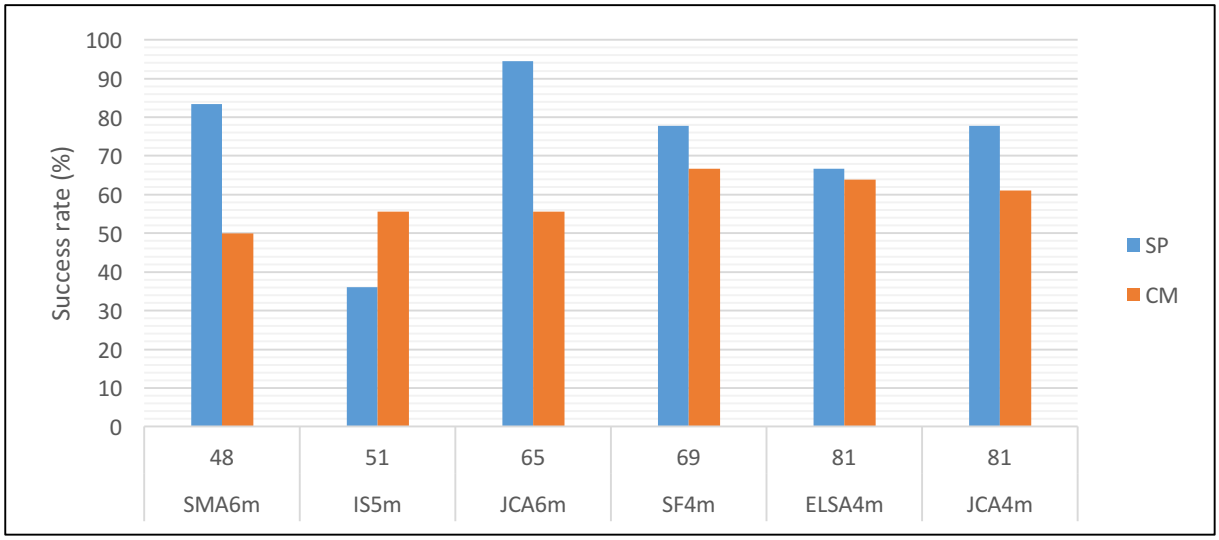


Figure 18. Mean success rates for temporally manipulated stimuli occurring both in speech block (SP) and cross-modal block (CM).

Figure 19 shows there were only four stimuli common for the non-speech and cross-modal block. In three out of four items, cross-modal instances have considerably lower success rate. Only JCA6m occurs with higher score in cross-modal block than in non-speech block.

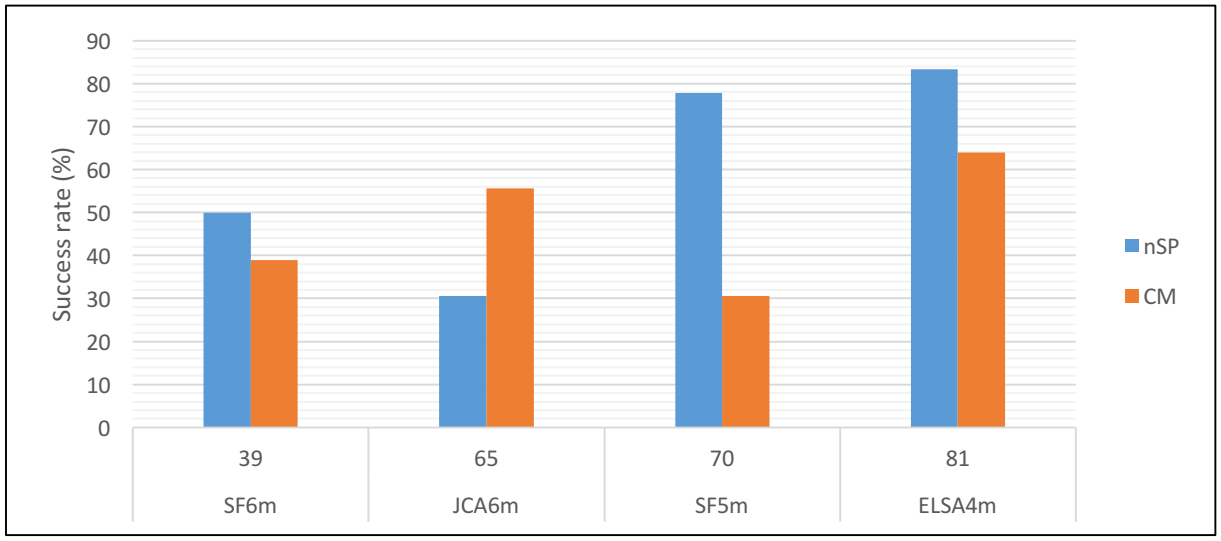


Figure 19. Mean success rates for temporally manipulated stimuli occurring both in non-speech block (nSP) and cross-modal block (CM).

Overall chart of mean success rate for modified items across different blocks is included in Appendix E.

4.9 Correlation with personal characteristics

To be able to correlate the performance in the experiment with the subjects' preceding musical training, fluency in foreign languages or current music-related activities, additional information was collected from the respondents. The group of eighteen subjects was designed to comprise respondents with practically no musical background who had never played any musical instrument, respondents who used to play an instrument but have little time to practise nowadays, and respondents who are in contact with their instruments and performing music on daily basis. Such assortment of respondents provided more or less proportionate sample of population and enabled us to trace their performance in the experiment depending on the length of their musical training, involvement in a band or choir, their current music activities, and their knowledge of English.

4.9.1 Music

First, the respondents were asked about the length of their formal or informal musical training. The duration was measured in years and the scale spanned twenty-two years. Figure 20 shows mean success rate for individual respondents aligned from the shortest to the longest period of musical training.

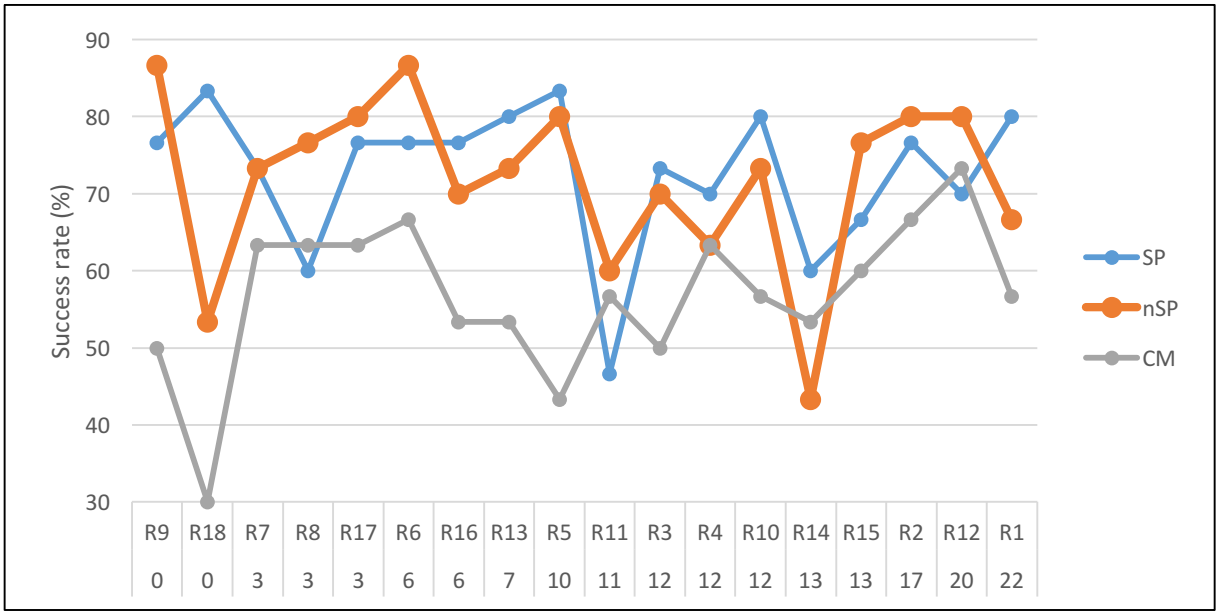


Figure 20. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) according to the increasing length of participants' involvement with music (in years).

The highest success rate was achieved for the non-speech block by subjects with zero (R9, 86.7%) and six (R6, 86.7%) years of musical training, and the same success rate of 80% was achieved by subjects with 10, 17 and 20 years of musical training. In the speech block, the highest success rate of 83.3% was achieved by a subject with zero years of musical education (R18).

4.9.2 Band / choir

However, the development of musical and rhythmic skills is particularly stimulated when the person is engaged in an activity that requires interaction with other people such as playing in a band or singing in a choir. Therefore, the next question that the respondents answered was whether they had had experience with being part of a musical collective, if any. The duration was measured in years and the scale spanned seventeen years. It is noteworthy that not infrequently did the length of respondents' musical training differ from the length of their involvement in a band or choir. Figure 21 shows mean success rate for individual respondents aligned from the shortest to the longest period of being part of a musical collective.

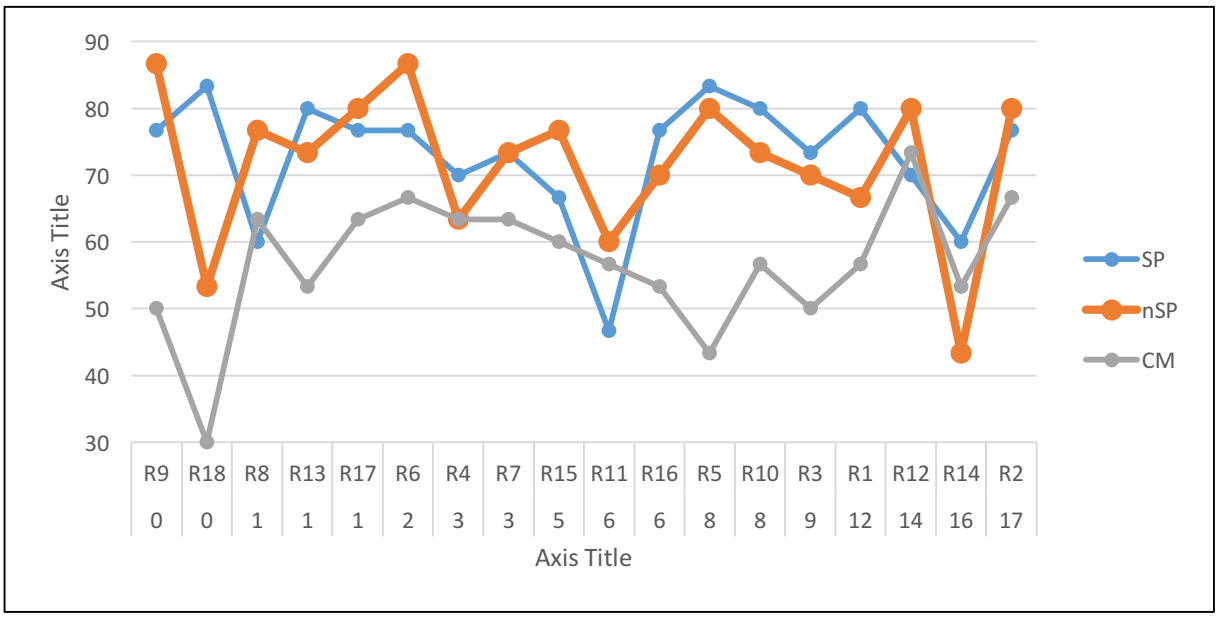


Figure 21. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) according to the increasing length of participants' involvement in a band / choir (in years).

Again, the highest success rates were surprisingly achieved by subjects with none or very short duration of their involvement in a musical collective. The highest score for the non-speech

block of 86.7% was achieved by the subjects R9 and R6 with experience of zero and two years of playing in a band, respectively. The same was evidenced for the speech block. The highest score of 83.3% was reached by R18, a subject with no collective musically-related experience. And in the cross-modal, no significant correlation existed between the length of the subjects' experience in musical collectives and their performance in the experiment.

4.9.3 Intensity of practice

Last music/related question concerned the subjects' current active involvement with music, which comprises only the activities when the subjects produce music with their musical instrument or voice. An average weekly amount of time was noted down for each subject. The time spent by casual listening to music without any direct interaction from the respondent was excluded. Figure 22 shows mean success rate for individual respondents aligned from the shortest to the longest period of weekly time they invest in the practise and production of music.

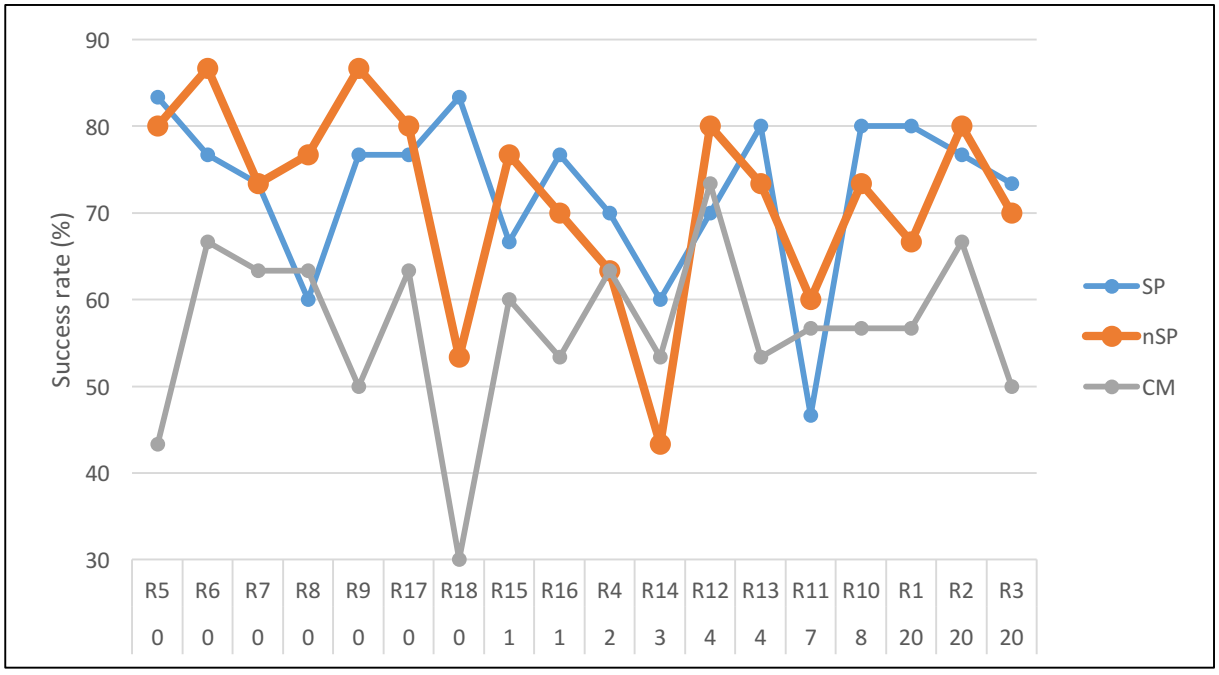


Figure 22. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) according to the increasing participants' intensity of practice (number of hours per week).

The highest success rates were achieved by subjects who reported that they were currently involved in no musically-related activities. For the non-speech block, the highest score of 86.7% was achieved by subjects R9 and R6, and for the speech block, the highest score of

83.3% was reached by R5 and R18.

4.9.4 Level of English

To examine the respondents' level of English proficiency on their overall performance in the experiment, information was collected about their current command of spoken English language according to CEFR. Majority of them were either on B1 (8 respondents) or B2 level (7 respondents). There was only one respondent with A2 English and two respondents with C1 English. The respondents were then categorized according to their proficiency level and the mean success rate of each CEFR group in individual blocks was plotted in Figure 23. Only in the speech block was the gradual increase apparent, starting on 66.67% for A2 level and reaching 81.67% for the speakers of C1. Neither for non-speech block nor for cross-modal block did the mean success rate correlate with CEFR level.

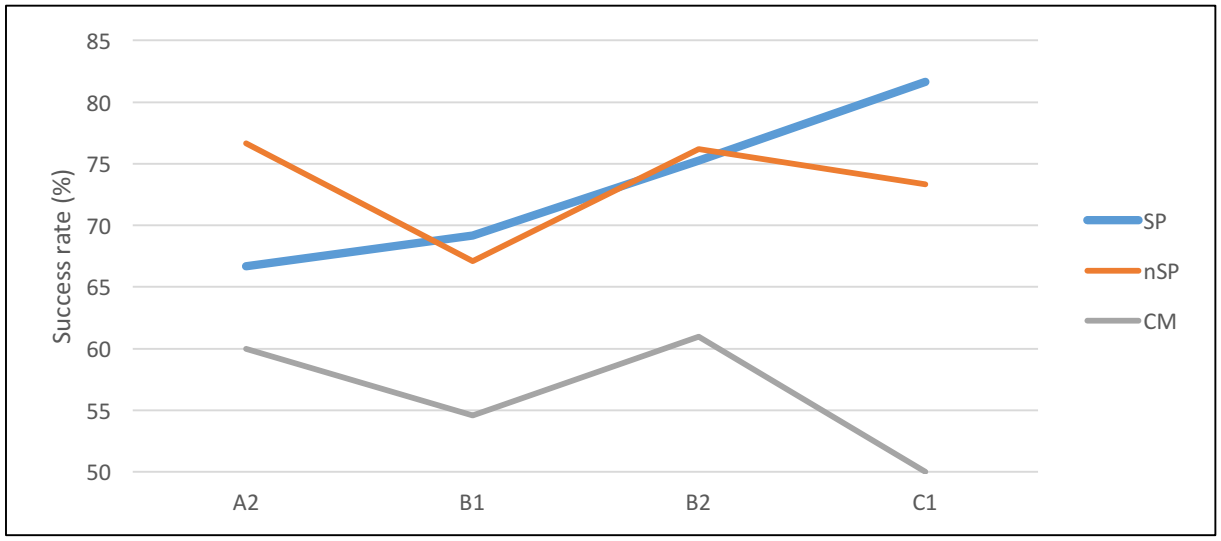


Figure 23. Mean success rates in the speech block (SP), non-speech block (nSP) and cross-modal block (CM) for participants of different L2 English level according to CEFR.

5. General discussion and conclusion

The overall success rate for all stimuli in the speech block and non-speech block was approximately the same. However, the cross-modal block proved to be considerably more difficult in several ways. The task of comparing two acoustic signals of different kind, one being speech signal and the following one rhythmic pattern performed by woodblock, was more demanding than comparing two stimuli of the same kind.

For each half of the respondents, the experiment was slightly altered in order to explore a potential influence of the first block on their performance. The respondents were more successful in the first two blocks when they were confronted initially with the speech block rather than non-speech block. This fact indicates that the memorization of speech phrases and their application to the corresponding, purely rhythmical patterns might be more straightforward than the other way around. As rhythm governs the flow of every spoken language, its abstraction from speech signal may be more natural than assigning English phrases to previously heard rhythmical phrases executed by metronome. However, for cross-modal block, the differences in success rates caused by first block influence were neutralized and much smaller.

The identification of manipulated stimuli was naturally harder than the identification of unmodified stimuli, and this fact was clearly reflected in the results too. Nevertheless, the subjects could not avoid incorrect classification even in the case when the phrases were rhythmically equal. False identification of unmodified items as rhythmically different was the most evident in the cross-modal block. In the speech and non-speech block, less than one fifth of unmodified stimuli were regarded as rhythmically unequal, but in the cross-modal block this number rose to as much as 42,59%. The uncertainty arose out of the task to compare two rhythmical phrases of different – speech and non-speech – types. In contrast to the unmodified stimuli, for the manipulated items the success rate was much more similar across all blocks. As for the manipulated stimuli, the cross-modal block did not represent a challenge for the respondents which would be considerably more difficult than the speech or non-speech block.

In each block, higher-than-average success rates for identification of rhythm deviations were achieved at different scales of modification. This duration was approximately 61 milliseconds for the non-speech block, 65 milliseconds for the speech block and 69 milliseconds for the cross-modal block (Q1). These findings indicate that the sensitivity to temporal modifications is higher for the non-speech stimuli than for speech stimuli. The fact that rhythmical phrases produced by metronome are easier to distinguish than English phrases was emphasized by respondents' impressionistic accounts during the experiment. The most commonly voiced concern was that speech contains load of unnecessary information that is more of an obstruction for reliable identification of its rhythm. Compared to human speech, the percussive phrases were professed to be 'pure' and the distractions that were present in the speech block were practically non-existent. H4 was thus contradicted by our results which show that the respondents could identify manipulation in shorter phrases better when they were of non-

speech, i.e. percussive, character. Common sense would suggest that the larger the scale of modification is, the more perceptually evident it should become. However, the results contradict this assumption and there are several factors that influenced them.

The results for items with manipulations applied to stressed syllables/beats were different than the modifications applied to unstressed positions. In the speech and cross-modal block, the discrimination task was easier when the manipulation was performed in the stressed syllables (Figure 7), which confirmed H1. Stressed syllables in speech are more prominent thanks to their longer duration, higher amplitude and higher pitch, which seems to have attracted the respondents' attention on the manipulated syllable better than on the unstressed, less prominent elements.

H3 assumed that the manipulations would be made more evident and more easily identifiable if a stressed syllable/beat succeeded the modified interval. However, this was confirmed only for the speech and cross-modal block. In the phrases in the speech block with the best results, the temporal modifications occur in unstressed syllables just before stressed syllables. In case of JCA6m, it's *the chief justice*, in SMA5m *molesting children*, and in SMA6m *in a rescue plan* (manipulated syllables underlined). The manipulation becomes more evident as the stressed syllables is more perceptually prominent than the unstressed one and it becomes easier to identify its dislocation from its original position.

The success rates for items in the speech and cross-modal block were higher when the manipulations were performed on a stressed syllable, or the syllable preceding a stressed syllable. On the contrary, the success rates for items in the non-speech block under these conditions were markedly lower. The identification of rhythmic deviations in the non-speech block was easier in unstressed beats. Subjects achieved better results in the non-speech block when the manipulations were performed on an unstressed beat, or the beat preceding an unstressed beat. These results for the non-speech block contradict our second and third hypothesis, achieving better results with unstressed rather than stressed beats in both positions.

Our findings suggest that the complexity expressed by the length of the phrase complicates successful identification of the modification (Q2). The success rate decreases with increasing number of syllables in the phrase, with the six-syllables-long phrases being the most difficult especially in the non-speech and cross-modal context. In the speech block, this distinction was weaker, which might indicate the improved ability of human ear to remember even longer

chunks of speech phrases more easily than those of their rhythmic counterparts.

Our results also indicate that the best results for the identification of manipulations were reached for the first syllable in the speech and cross-modal block, followed by the third syllable. On the other hand, it was the second beat which reached the highest rate in the non-speech block. The results are consistent with H5 as the modification in the fourth and fifth syllables/beats proved to be the most difficult, with the success rate decreasing from 50% to as low as 25%.

Mean intra-respondent consistency reached approximately 70% in the experiment, with the lowest variability in the speech block, and the highest variability in the cross-modal block. Intra-respondent consistency clearly deteriorates together with the increasing difficulty of the discrimination task in individual blocks. Comparisons of the first and second occurrences of testing and controlling items in the blocks revealed that for the majority of items, the success rate had declined considerably for the second occurrence. The worst results were achieved for the speech block with an average -10% decrease in the success rate, and the lowest decrease of -3,89% occurred in the cross-modal block. These results suggest that the best self-correction occurred in the cross-modal block, where four items had higher success rate in their second occurrence. There were three such items for the non-speech block, and only two for the speech block.

The analysis of re-occurrences of individual temporally manipulated stimuli in different speech blocks revealed that in the majority of the cases, the identification of a speech item was more successful than of its non-speech and cross-modal counterpart.

None of the personal variables for the length of music education, involvement in a band or choir or the amount of current musical activities showed any significant correlation with the ability to distinguish rhythmical deviations in the experiment. H1 could not be confirmed and it was also contradicted by the fact that participants with zero years of preceding musical training achieved the highest success rate for both speech and non-speech block.

Influence of the respondents' proficiency in spoken English on results was evident only in the speech block. Successful identification of rhythm manipulations in short English phrases was gradually improving alongside rising command of the language. For the other two blocks, non-speech and cross-modal, the correlation was not detected.

6. Conclusion

The present study was designed to examine the ability to hear differences in the realm of speech and non-speech rhythm. Fifteen recordings of short English phrases with their temporally manipulated counterparts were carefully synthetized in music software to create rhythmically identical non-speech phrases using percussive sounds. The stimuli were arranged into three blocks of thirty items. Each block contained twenty unique items and ten control items, used later for the calculations of intra-respondent consistency. The speech and non-speech block focused on the ability hear slight rhythmical differences in speech and non-speech, i.e. music, and the cross-modal block explored the hypothesis of interconnectedness of these two domains.

In the speech block, the identification of rhythmic deviations proved the least difficult. The best results were achieved when the manipulations occurred in the 1st or 3rd syllable of the phrase, and when this syllable either preceded a stressed syllable or it was stressed itself. Intra-respondent consistency was the highest for the speech block, too.

The best results for the non-speech items were achieved when the manipulations occurred in the 2nd beat of the rhythmic phrase, and when this beat either preceded an unstressed beat or it was unstressed itself. This finding acts contrary to the speech block, where the stressed syllables or the following stressed syllables were the major aid for better identification of rhythm deviations.

For each block, the boundaries for reliable identification of rhythm deviations occurred at different scales of modification. Higher-than-average sensitivity rate for rhythmically manipulated recordings was found for non-speech items with 61 milliseconds long deviations already distinguishable, compared to 65 ms in the speech and 69 ms in the cross-modal block.

The hypothesis that the cross-modal block will represent the most demanding task for the subjects was confirmed. Comparing two phrases of different type was the most difficult part of the experiment for the subjects, who expressed high level of uncertainty in assessing whether the rhythmical representations corresponded to its speech counterparts. This was also underlined by the fact that the incorrect classification of identical stimuli was the most common for the cross-modal block. Although placement of the beat in the experiment followed widely accepted convention that P-centers occur at or very near the vowel onset (Cummins & Port, 1998; Fowler, 1983; Scott, 1993; Scott, 1998), weak performance in the cross-modal block

challenges the concept of universal location of P-centers. Our findings seem to point to the variability of P-center placement which may be unique to each phrase due to other variables, such as prevocalic elements, duration of the vowel or syllable, or the spectral envelope (Fox & Lehiste, 1987; Scott, 1993; Harsin, 1997; Howell, 1988; Pompino-Marschall, 1989).

The influence of the first block was evident, too. When the subjects were confronted with the speech block first, their performance in the experiment improved. Universal familiarity with speech seemed to provide better initial confidence to the subjects in their evaluations than purely rhythmical patterns.

It was also shown that there is a limited capacity of human ear to memorize rhythmical patterns, and with the increasing length of the phrase this ability deteriorates. In our study this is indicated by a significant drop in the success rate for the phrases of six syllables/beats, compared to the phrases with the length of four or five syllables/beats.

The hypothesis that musically trained or musically active subjects will achieve better results wasn't confirmed. Neither the preceding musical training, involvement in a musical collective, nor the amount of time currently devoted to music-related activities influenced the results. Paradoxically, it was often the subjects who never picked up an instrument and never played in a band that achieved the best results. The only interference with the results was the English proficiency of the respondents. The better their English L2 level was, the higher success rates they achieved in the speech block.

These findings are related to the language used in the phrases that were used for the experiment and agree with those of Lidji et al. (2011) that long-term linguistic experience with a stress-timed language can differentiate speakers' entrainment to rhythmic regularities. Subjects with better command of and longer exposure to English language are assumed to more accustomed to the natural flow of the language. This ability may have enabled participants to hear the rhythmic deviations more readily than the subjects with lower level of spoken L2 English. The two other blocks were unaffected by the participants' level of L2 English.

The present diploma thesis attempted to explore two different kinds of short rhythmical phrases and the ability of musically trained and untrained population to discover rhythm manipulations in different positions and of different lengths. To analyse the specific segmental features of speech phrases that facilitate identification of rhythm deviations in detail was beyond the limits of the current diploma thesis and further research is recommended.

Bibliography

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Abraham, G. (1974). *The tradition of Western music* (pp. 62–83). Berkeley: University of California Press.

Adams, C. (1979). *English Speech Rhythm and the Foreign Learner*. The Hague: Mouton.

Allen, G. D. (1972). "The location of rhythmic stress beats in English." *Language and Speech*, 72-100; 179-195.

Arvaniti, A. (2012). "The usefulness of metrics in the quantification of speech rhythm." *Journal of Phonetics*, 40, 351-373.

Atkinson, K. (1968). "Language Identification from Nonsegmental Cues." *The Journal of the Acoustical Society of America*, 44, 378.

Bahrick, L. E., & Pickens, J. N. (1988). "Classification of bimodal English and Spanish language passages by infants," *Infant Behavior and Development*, 11, 277- 296.

Bailey J. A., Penhune V. B. (2010). "Rhythm synchronization performance and auditory working memory in early- and late-trained musicians." *Experimental Brain Research*, 204, 91–101.

Barbosa, P. A., et al. (2005). "Abstractness in Speech-Metronome Synchronisation: P-centers as Cyclic Attractors." *Ninth European Conference on Speech Communication and Technology*, 1441-4.

Barry, W. J., et al. (2003). "Do rhythm measures tell us anything about language type?" *Proceedings of ICPhS*. Barcelona, Spain.

Barry, W. J., (2007). "Rhythm as an L2 Problem: How prosodic is it?" In J. Trouvain & U. Gut (Eds.): *Non-Native Prosody – Phonetic Description and Teaching Practice.* Berlin: Mouton. pp. 97-120.

Benadon, F. (2013). "Metrical perception of trisyllabic speech rhythms." *Psychological Research*, 78/1, pp. 113–123.

Bengtsson, S. L., et al. (2005). "Extensive piano practicing has regionally specific effects on white matter development." *Nature Neuroscience*, 8, 1148–1150.

Bermudez, P., et al. (2009). "Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry." *Cerebral Cortex,* 19, 1583–1596.

Bertrán, A. P. (1999). "Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages." *Language Design*, 2, 103-130.

Besson, M., et al. (2007). "Influence of musical expertise and musical training on pitch processing in music and language." *Restorative Neurology and Neuroscience*, 25, 399–410.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer, version 5.3.11., www.praat.org.

Bolinger, D. L. (1958). "A theory of pitch accent in English," *Word*, 14, 109–119.

Bonte, R. (1975). "Can you identify a language by its prosody?" Unpublished MA thesis. University of California, Berkeley.

Bosch, L. & Sebastián-Gallés, N. (1997). "Native language recognition abilities in 4-month-old infants from monolingual and bilingual environments." *Cognition,* 65, 33-69.

Boucourechliev, A. (1993). *Le Langage Musical*. Collections Les Chemins de la Musique. Paris: Fayard.

Brown, S. (2000). "The musilanguage model of music evolution," in *The Origin of Music*, Eds. Wallin N. L., Merker B., & Brown S. Cambridge, MA: MIT Press. pp. 271–300.

Brown, W. (2013). *Time in English Verse Rhythm*. London: Forgotten Books.

Bush, N. C. (1967). "Some Acoustic Parameters of Speech and Their Relationships to the Perception of Dialect Differences." *TESOL Quarterly*, 1/3, pp. 20-30.

Carter, P. M. (2005). "Quantifying Rhythmic Differences between Spanish, English, and Hispanic English." In R. Gees, ed., Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages. Amsterdam: John Benjamins. 63–75.

Cenoz, J. & Lecumberri L. G. (1999). "The acquisition of English pronunciation: learner's views." *International Journal of Applied Linguistics*, 9, 3-15.

Chela-Flores, B. (1994). "On the acquisition of English rhythm: theoretical and practical issues." *International Review of Applied Linguistics,* 32, 232–42.

Classe, A. (1939). *The Rhythm of English Prose*. Oxford: Basil Blackwell.

Coleman, C. (1974). "A study of acoustical and perceptual attributes of isochrony in spoken English." Unpublished doctoral dissertation. Washington: University of Washington Press.

Cooper, A. M., et al. (1986). "P-centers are unaffected by phonetic categorization." *Perception & Psychophysics*, 39, 187–196.

Cooper, G. & Meyer, L. B. (1960). *The rhythmic structure of music*. Chicago: University of Chicago Press.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (1996). "The past, present and future of English rhythm." *Speak Out, Newsletter of the IATEFL Pronunciation Special Interest Group,* 18, 8-13.

Crystal, D. (1997). *The Cambridge encyclopaedia of language*. Cambridge: Cambridge University Press.

Cummins, F. (2009). "Rhythm as an affordance for the entrainment of movement." *Phonetica*, 66, 15–28.

Cummins, F. & Port, R. (1998). "Rhythmic constraints on stress timing in English." *Journal of Phonetics*, 26, 145–171.

Darwin, C. (1871/1981). "The descent of man and selection in relation to sex," Princeton NJ: Princeton University Press.

Dasher, R. & Bolinger, D. (1982). "On pre-accentual lengthening." *Journal of the International Phonetic Association*, 12, 58-69.

Dauer, R. M. (1983). "Stress-timing and syllable-timing reanalyzed." *Journal of Phonetics*, 11, 51–62.

Dauer, R. M. (1987). “Phonetic and phonological components of language rhythm.” *The 11th International Congress of Phonetic Sciences*, vol. 5, 447-450.

de Jong, K. J. (1992). “Acoustic and Articulatory Correlates of P-center Perception.” *UCLA Working Papers in Phonetics*, 81, 66 - 75.

de Jong, K. J. (1994). “The Correlation of P-center Adjustments with Articulatory and Acoustic Events.” *Perception and Psychophysics*, 56, 447 - 460.

de Jong, K.J. (2001). “Effects of Syllable Affiliation and Consonant Voicing on Temporal Adjustment in a Repetitive Speech Production Task.” *Journal of Speech, Language, and Hearing Research*, 44, 826-840.

de Pijper, J. R. (1983). “Modelling British English Intonation: An Analysis by Resynthesis of British English Intonation.” Foris Publications, USA.

Dehaene-Lambertz, G. & Houston, D. (1998). “Faster orientation latencies toward native language in two-month old infants,” *Language Speech,* 41, 21-43.

Dellwo, et al. (2005). “Influence of L1 on rhythm L2.” Workshop at Universität des Saarlandes, Saarbrücken.

Dellwo, V. & Wagner, P. (2003). “Relations between language rhythm and speech rate.” In *Proceedings of the15th international congress of phonetics sciences*, 471–474. Barcelona, Spain.

Drake, C. (1993). “Reproduction of musical rhythms by children, adult musicians, and adult nonmusicians.” *Perception and Psychophysics*, 53, 25–33.

Eds. May, M. & Wisse, J. (2001). Cicero: *On the Ideal Orator*. New York: Oxford University Press.

Elbert, T., et al. (1995). “Increased cortical representation of the fingers of the left hand in string players.” *Science,* 270, 305–307.

Essens, P. J., & Povel, D.-J. (1985). “Metrical and nonmetrical representations of temporal patterns.” *Perception de Psychophysics*, 37, 1-7.

Falk, D. (2004). “Prelinguistic evolution in early hominins: whence motherese?” *Behavioral and Brain Sciences*, 27, 491–503.

Fear, B. B., at al. (1995). “The strong/weak syllable distinction in English.” *Journal of the Acoustical Society of America*, 97/3, 1893-1904.

Fitch W. T. (2010). *The Evolution of Language*. New York: Cambridge University Press.

Fowler, C. A. (1979). “Perceptual centers in speech production and perception.” Perception & Psycophysics, 25, 375–388.

Fowler, C. A. (1983). “Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet.” *Journal of Experimental Psychology*, 112, 386–412.

Fox, R. A. & Lehiste, I. (1987). “The effect of vowel quality variations on stress-beat location.” *Journal of Phonetics*, 15, 1–13.

Fraisse, P. (1982). “Rhythm and tempo.” *Psychology of Music,* 1, 149–180.

Fraisse, P. (1974). *Psychologie du rythme*. Paris: PUF.

Frost, D. (2011). “Stress and cues to relative prominence in English and French: A perceptual study.” *Journal of the International Phonetic Association*, 41/1, 67-84.

Fry, D. B. (1965). "The dependence of stress judgments on vowel formant structure," *Proceedings of the 5th International Congress of Phonetics Sciences*, 306–311.

Fry, D. B. (1958). "Experiments in the perception of stress." *Language Speech,* 1, 126–153.

Fry, D. B. (1955). "Duration and intensity as physical correlates of linguistic stress." *Journal of the Acoustical Society of America*, 27, 765.

Gaser, C., & Schlaug, G. (2003). "Brain structures differ between musicians and non-musicians." Journal of Neuroscience, 23, 9240–9245.

Gass, S. M. & Selinker, L. (2008). *Second Language Acquisition: an Introductory Course*. New York and London: Routledge.

Ghitza, O. & Greenberg, S. (2009). "On the possible role of brain rhythms in speech perception." *Phonetica*, 66, 113-126

Gibbon, D., & Gut, U. (2001). "Measuring speech rhythm." *Proceedings of Eurospeech*, 91–94. Aalborg.

Gimson, A. C. (1975). *A practical Course of English Pronunciation: A Perceptual Approach*. London: Edward Arnold.

Gordon, R. L., Magne, C. L. & Large, E. W. (2011). "EEG correlates of song prosody: A new look at the relationship between linguistic and musical rhythm." *Frontiers in Psychology*, 2, 352.

Grabe, E. & Low, E. L. (2002). "Durational variability in speech and the rhythm class hypothesis." In N.Warner, & C. Gussenhoven (Eds.). *Papers in laboratory phonology 7*. Berlin: Mouton de Gruyter.

Grahn, J. A. (2012). "Neural mechanisms of rhythm perception: current findings and future perspectives." *Topics in Cognitive Science*, 4, 585–606

Grosser, W. (1993), "Aspects of intonation L2 acquisition", in B. Kettemann & W. Wieden (eds), *Current Issues in European Second Language Acquisition Research*. Tübingen: Gunter Narr Verlag, 81-94.

Guilbault, C. (2002). "The Acquisition of French Rhythm by Second Language Learners." PhD thesis, University of Alberta.

Gut, U. (2011). "Rhythm in L2 Speech." University of Munster, Germany.

Gut, U. (2003). "Non-native speech rhythm in German." *Proceedings of 15th International Congress of Phonetic Sciences*, Barcelona. 2437–2440.

Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. Bradford, MIT Press.

Hannon, E. E., & Trainor, L. J. (2007). "Music acquisition: Effects of enculturation and formal training on development." *Trends in Cognitive Sciences*, 11, 466–72.

Harsin, C. A. (1997). "Perceptual-center modeling is affected by including acoustic rate-of-change modulations." *Perception & Psychophysics*, 59, 243–251.

Hausen, M., et al. (2013). "Music and speech prosody: A common rhythm." *Frontiers in Psychology*, 4, 566.

Hinton, S. C., & Rauscher, F. H. (2003). "Type of music training selectively influences perceptual processing." In R. Kopiez, A. C. Lehmann, I. Wolther & C. Wolf (Eds), *Proceedings of the 5th Triennial ESCOM conference*, 89–92. Hanover: Germany.

Hoequist, C. (1983). "Syllable duration in stress-, syllable-, and mora-timed languages." *Phonetica,* 40, 203-237.

Howell, P. (1988). "Prediction of P-centre location from the distribution of energy in the amplitude envelope." *Perception and Psychophysics*, 43, 90–93.

Hutchinson, S., et al. (2003). "Cerebellar volume: gender and musicianship effects." *Cerebral Cortex,* 13, 943–949.

Jackendoff, R. (1989). "A comparison of rhythmic structures in music and language," In *Rhythm and Meter*, Eds. Kiparsky P., Youmans G. San Diego: Academic Press. 15–44.

Jenkins, R. A. (1961). "Perception of Pitch, Timbre, and Loudness." *The Journal of the Acoustical Society of America,* 33.

Jones, M. R. (1987). "Dynamic pattern structure in music: Recent theory and research." *Perception de Psychophysics*, 41, 621-634.

Jones, M. R. (1990). "Learning and the development of expectancies: An interactionist approach." *Psychomusicology*, 9, 193-228.

Jusczyk, et al. (1993). "Infants' sensitivity to the sound pattern of native language words," *Journal of Memory and Language,* 32, 402-420.

Jusczyk, R., & Krumhansl, C. (1993). "Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure." *Journal of Experimental Psychology: Human Perception and Performance*, 19/3, 627–640

Juslin, P. N. & Laukka P. (2003). "Communication of emotions in vocal expression and music performance: different channels, same code?" *Psychological Bulletin*, 129, 770–814.

Knowles, G. (1995). "Review of Approaches to Pronunciation Teaching." *ELT Journal,* 49, 286–9.

Kohler, K. (2009a). "Whither speech rhythm research?" *Phonetica*, 66, 5-14.

Kohler, K. (2009b). "Rhythm in speech and language. A new research paradigm." *Phonetica*, 66, 29-45.

Kotilahti K., et al. (2010). "Hemodynamic responses to speech and music in newborn infants." *Human Brain Mapping,* 31, 595–603.

Kraus, N. et al. (2009). "Experience-induced malleability in neural encoding of pitch, timbre, and timing." *Annals of the New York Academy of Science*s, 1169, 543–557.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Lai, C., et al. (2013). "Applying rhythm metrics to non-native spontaneous speech." *Proceedings of SLaTE*. Grenoble: France.

Large, E. W. (2008). "Resonating to musical rhythm: Theory and experiment." In S. Grondin (Ed.), *Psychology of time*. Bingley, UK: Emerald.

Lehiste, I. (1973). "Rhythmic units and syntactic units in production and perception." *The Journal of the Acoustical Society of America*, 54/5, 1228–1234.

Lehiste, I. (1977). "Isochrony reconsidered." *Journal of Phonetics*, 5, 253–263.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lidji, P., et al. (2011). "Listeners feel the beat: Entrainment to English and French speech rhythms." *Psychonomic Bulletin & Review*, 18/6, 1035–1041.

Lieberman, P. (1960). "Some acoustic correlates of word stress in American English." *Journal of the Acoustical Society of America*, 32, 451–454.

Lin, H. & Wang, Q. (2005). "Vowel quantity and consonant variance: A comparison between Chinese and English." *Proceedings of Between Stress and Tone*. Leiden: Belgium.

Lindblom, B. (1970). "Temporal Organization of Syllabic Processes." *Paper at the 79th ASA meeting.* Atlantic City.

Lloyd James, A. (1940*). Speech Signals in Telephony*. London: Pitman & Sons.

Low, E. L., Grabe, E., & Nolan, F. (2000). "Quantitative characterisations of speech rhythm: 'Syllabletiming' in Singapore English." *Language and Speech*, 43, 377–401.

Machač, P. & Skarnitzl, R. (2009). *Principles of Phonetic Segmentation*. Praha: Epocha.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Maidment, J. A. (1976). "Voice fundamental frequency characteristics as language differentiators." *Speech and Hearing: Work in Progress*, 2, 74–93.

Maidment, J. A. (1983). "Language recognition and prosody: Further evidence." *Speech, Hearing and Language: Work in Progress*, 1, 133–141.

Marcus, S. M. (1981). "Acoustic determinants of Perceptual center (P-center) location." *Perception and Psychophysics*, 30, 247-256.

Mehler, J., et al. (1986). "Discrimination de la langue maternelle par le nouveau-né." *Comptesrendus de l'Académie des Sciences de Paris*, 303/III, 637-640.

Mehler, J., et al. (1988). "A precursor of language acquisition in young infants." *Cognition,* 29, 143-178.

Milovanov, R., & Tervaniemi, M. (2011). "The interplay between musical and linguistic aptitudes: A review." *Frontiers in Psychology*, 2, 321.

Mithen, S. (2005). "The singing Neanderthals: The origins of music, language, mind and body." London: Weidenfeld & Nicholson.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). "Two-day-olds prefer their native language." *Infant Behavior and Development*, 16, 495-500.

Morton J, & Jassem W. (1965). "Acoustic correlates of stress." *Language and Speech*, 8, 159–181

Morton, J., Marcus, S., & Frankish, C. (1976). "Perceptual centers (P-centers)." *Psychological Review*, 83, 405-408.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). "Language discrimination by English learning 5 month olds: Effects of rhythm and familiarity." *Journal of Memory and Language*, 43, 1–19.

Nooteboom, S. G. (1997). "The prosody of speech: Melody and rhythm." In W.J. Hardcastle & J. Laver (Eds.), *Handbook of Phonetic Sciences,* 640-673. Oxford: Blackwell.

Ohala, J. J. & Gilbert, J. B. (1979). “Listeners' ability to identify languages by their prosody.” In: Léon, P. and Rossi, M. [Ed], Problémes de prosodie, Vol. II. Ottawa, Didier, p. 123-131.

Pantev C, et al. (1998). “Increased auditory cortical representation in musicians.” *Nature*, 392, 811–814.

Patel A. D. (1998). “Syntactic processing in language and music: different cognitive operations, similar neural resources?” *Music Perception*, 16, 27–42.

Patel A. D. (2003). “Language, music, syntax and the brain.” *Nature Neuroscience*, 6, 674–681.

Patel A. D. (2003b). “Rhythm in language and music: parallels and differences.” *Annals of the New York Academy of Science*s, 999, 140–143.

Patel A. D. (2008). “Music, Language and the Brain.” New York: Oxford University Press.

Patel, A., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, 87, 35-45.

Patel, A. D., Löfqvist, A. & Naito, W. 1999. “The acoustics and kinematics of regularly timed speech: A database and method for the study of the P-center problem.” *Proceedings of the 14th International Congress of Phonetic Sciences,* 405–408.

Pennigton, M. C. (1996). *Phonology in English Language Teaching: An International Approach*. Longman.

Peter, V., McArthur, G., & Thompson, W. F. (2012). “Discrimination of stress in speech and music: A mismatch negativity (MMN) study.” *Psychophysiology*, 49/12, 1590–1600

Pike, K. L. (1945). *The Intonation of American English*. University of Michigan Press: Ann Arbor.

Pinker, S. (1997). *How the Mind Works*. New York, NY: W. W. Norton & Company.

Piske, T. et al. (2000). “Factors affecting degree of foreign accent in an L2: a review.” *Journal of Phonetics*, 29, 191-215

Plag I., et al. (2011). “Acoustic correlates of primary and secondary stress in north American English.” *Journal of Phonetics*, 39, 362–374.

Pompino-Marschall, B. (1989). “On the psychoacoustic nature of the P-centre phenomenon.” *Journal of Phonetics*, 17, 175–192.

Port, R. (2003). “Meter and speech.” *Journal of Phonetics*, 31, 599–611.

Povel, D. J. & Essens, P. (1985). “Perception of Temporal Patterns” *Music Perception*, 2/4, 411-440.

Povel, D. J. (1981). “Internal representation of simple temporal patterns.” *Journal of Experimental Psychology: Human Perception and Performance*, 7, 3–18.

Ramus, F. (2002). “Language discrimination by newborns: teasing apart phonotactic, rhythmic and intonational cues.” *Annual Review of Language Acquisition*, 2, 85-115.

Ramus, F., et al. (1999). “Correlates of linguistic rhythm.” *Cognition,* 73, 265±292.

Ramus, F., et al. (2003). “The psychological reality of rhythm classes.” *Proceedings of ICPhS 2003*. Barcelona, Spain.

Ramus, F., et al. (2000). “Language discrimination by human newborns and by cotton-top tamarin monkeys.” *Science*, 288/5464, 349–351.

Rapp, K. (1971). “A study of syllable timing.” *Papers from the Institute of Linguistics*, 8, 14–19.

Repp, B. H. (1999). “Detecting deviations from metronomic timing in music: Effects of perceptual structure on the mental timekeeper.” *Perception & Psychophysics*, 61, 529–548.

Richardson, J. C. (1973). “The identification by voice of speaker belonging to two ethnic groups.” Unpub. PhD. Dissertation. Ohio State University.

Roach, P. (1982). “On the distinction between stress-timed and syllable-timed languages.” Originally published in *Linguistic Controversies*, ed. D. Crystal, 1982, 73-79.

Roach, P. (1991). *English Phonetics and Phonology*. Cambridge: Cambridge University Press.

Roncaglia-Denissen, M., et al. (2016). “The enhanced musical rhythmic perception in second language learners.” *Frontiers in Human Neuroscience*, 10, 288.

Rousseau, J. J. (1781/1993). *Essai Sur L'origine des Langues*. Paris: Flammarion.

Schellenberg, E. G. (2005). “Music and cognitive abilities.” *Current Directions in Psychological Science*, 14/6, 317–320.

Schlaug, G., et al. (1995a). “Increased corpus callosum size in musicians.” *Neuropsychologia,* 33, 1047–1055.

Schmithorst, V. J., & Wilke, M. (2002). “Differences in white matter architecture between musicians and nonmusicians: a diffusion tensor imaging study.” *Neuroscience Letters*, 321, 57–60.

Scott, S. K. (1993). “P-centres in speech: An acoustic analysis.” Unpublished doctoral dissertation, University College, London.

Scott, S. K. (1998). “The point of P-centres.” *Psychological Research*, 61, 4–11.

Skarnitzl, R. (2010). “Prague Phonetic Corpus: status report.” *Phonetica Pragensia*, 12/1, 65-67.

Sluijter, A. M. C. & Van Heuven, V. J. (1996). “Spectral balance as an acoustic correlate of linguistic stress.” *Journal of the Acoustical Society of America*, 100/4, 2471-2485.

Spencer H. (1857). “The origin and function of music.” *Fraser's Magazine*, 56, 396–408.

Stockmal, V., Markus, D. & Bond, D. (2005). “Measures of Native and Non-Native Rhythm in a Quantity Language.” *Language and Speech,* 48, 55–63.

Stoffer, T. H. (1985). “Representation of phrase structure in the perception of music.” *Music Perception,* 3, 191-220.

Swain, J. P. (1997). “Musical languages.” New York: W.W. Norton.

Tajima, K. & Port, R. (2003). “Speech rhythm in English and Japanese.” In J. Local, R. Ogden, & R. Temple (Eds.), P*apers in Laboratory Phonology VI*. Cambridge: Cambridge University Press

Trehub, E.S. (2003). “The Developmental origins of musicality.” *Nature Neuroscience*, 6/7, 669-673;

Tzounopoulos, T. & Kraus N. (2009). “Learning to encode timing: Mechanisms of plasticity in the auditory brainstem.” *Neuron*, 62, 463–469.

van der Hulst, H. (2010). *A Survey of Word Accentual Patterns in the Languages of the World*. The Hague: Mouton.

Villing, et al. (2011). “Measuring perceptual centers using the phase correction response.” *Attention, Perception, & Psychophysics,* 73, 1614–1629

Volín, J (2010). “On the significance of the temporal structuring of speech.” In: Markéta Malá — Pavlína Šaldová (eds.), *...for thy speech bewrayeth thee* (A Festschrift for Libuše Dušková). Praha: Filozofická fakulta Univerzity Karlovy v Praze, 289–305.

Volín, J. (2002). *IPA-Based Transcription for Czech Students of English*. Praha: Karolinum.

Volín, J. (2007). *Statistické Metody ve Fonetickém Výzkumu*. Praha: Epocha.

Warner, N. & Arai, T. (2001). “Japanese mora-timing: a review.” *Phonetica,* 58, 1–25.

Wenk, B. J. (1987). “Just in time: on speech rhythms in music.” *Linguistics*, 25, 969–981.

White, L. & Mattys, S. L. (2007a). “Rhythmic typology and variation in first and second languages.” In P. Prieto, J. Mascaró , & M.-J. Solé (Eds.), *Segmental and prosodic issues in romance phonology. Current issues in linguistic theory series*. Amsterdam, Philadelphia: John Benjamins.

White, L. & Mattys, S. L. (2007b). “Calibrating rhythm: First language and second language studies.” *Journal of Phonetics*, 35, 501-522.

White, L. & Mattys, S. L. (2008). “That elusive rhythm. Pros and cons of rhythm metrics.” Laboratory Phonology, 11.

Whitworth, N. (2002). “Speech rhythm production in three German-English bilingual families.” *Leeds Working Papers in Linguistics and Phonetics.*

Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris.

Xu, Yi & Ching Xu. (2005). “Phonetic realization of focus in English declarative intonation.” *Journal of Phonetics,* 33, 159–197.

Yeston, M. (1976). *The stratification of musical rhythm*. New Haven: Yale University Press.

Appendix

APPENDIX A

The order of stimuli in the speech block with randomly assigned Same/Different values and alternating length of four, five and six syllables.

No.	Item	S/D	NoSyll
1	ELSA4	S	4
2	IS5m	D	5
3	JCA6m	D	6
4	SF4	S	4
5	SMA5m	D	5
6	ELSA6	S	6
7	IS4	S	4
8	JCA5m	D	5
9	SF6	S	6
10	SMA4m	D	4
11	ELSA5m	D	5
12	IS6m	D	6
13	JCA4m	D	4
14	SF5	S	5
15	SMA6m	D	6
16	ELSA4m	D	4
17	IS5	S	5
18	JCA6	S	6
19	SF4m	D	4
20	SMA5	S	5
21	IS5m	D	5
22	SF4	S	4
23	ELSA6	S	6
24	JCA5m	D	5
25	SMA4m	D	4
26	IS6m	D	6
27	SF5	S	5
28	ELSA4m	D	4
29	JCA6	S	6
30	SMA5	S	5

The order of stimuli in the non-speech block with randomly assigned Same/Different values and alternating length of four, five and six syllables.

No.	Item	S/D	NoSyll
1	ELSA4	S	4
2	IS5m	D	5
3	JCA6m	D	6
4	SF4	S	4
5	SMA5m	D	5
6	ELSA6	S	6
7	IS4	S	4
8	JCA5m	D	5
9	SF6	S	6
10	SMA4m	D	4
11	ELSA5m	D	5
12	IS6m	D	6
13	JCA4m	D	4
14	SF5	S	5
15	SMA6m	D	6
16	ELSA4m	D	4
17	IS5	S	5
18	JCA6	S	6
19	SF4m	D	4
20	SMA5	S	5
21	IS5m	D	5
22	SF4	S	4
23	ELSA6	S	6
24	JCA5m	D	5
25	SMA4m	D	4
26	IS6m	D	6
27	SF5	S	5
28	ELSA4m	D	4
29	JCA6	S	6
30	SMA5	S	5

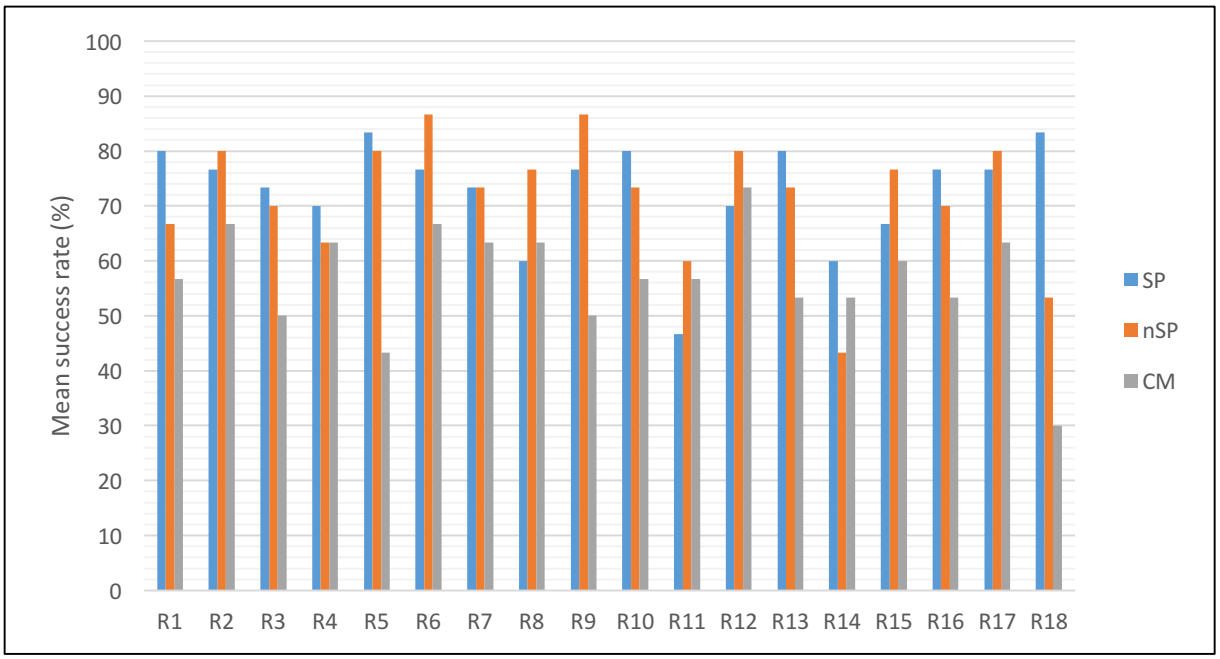
The order of stimuli in the cross-modal block with randomly assigned Same/Different values and alternating length of four, five and six syllables.

No.	Item	S/D	NoSyll
1	JCA4m	D	4
2	SF5m	D	5
3	SMA6	S	6
4	ELSA4m	D	4
5	IS5m	D	5
6	JCA6	S	6
7	SF4m	D	4

8	SMA5	S	5
9	ELSA6	S	6
10	IS4	S	4
11	JCA5	S	5
12	SF6m	D	6
13	SMA4m	D	4
14	ELSA5	S	5
15	IS6	S	6
16	JCA4	S	4
17	SF5	S	5
18	SMA6m	D	6
19	ELSA4	S	4
20	IS5	S	5
21	SF5m	D	5
22	ELSA4m	D	4
23	JCA6	S	6
24	SMA5	S	5
25	IS4	S	4
26	SF6m	D	6
27	ELSA5	S	5
28	JCA4	S	4
29	SMA6m	D	6
30	IS5	S	5

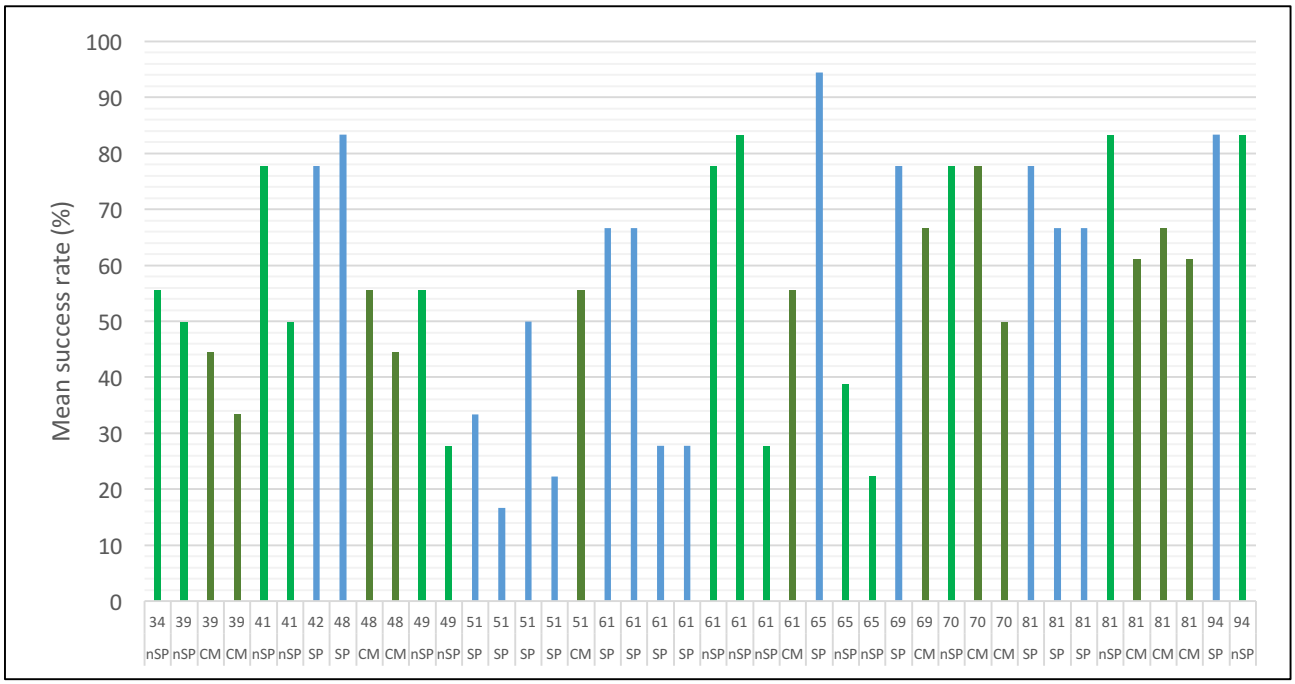
APPENDIX B

Mean success rate for all respondents (R1 – R18) in speech block (SP), non-speech block (nSP) and cross-modal block (CM).



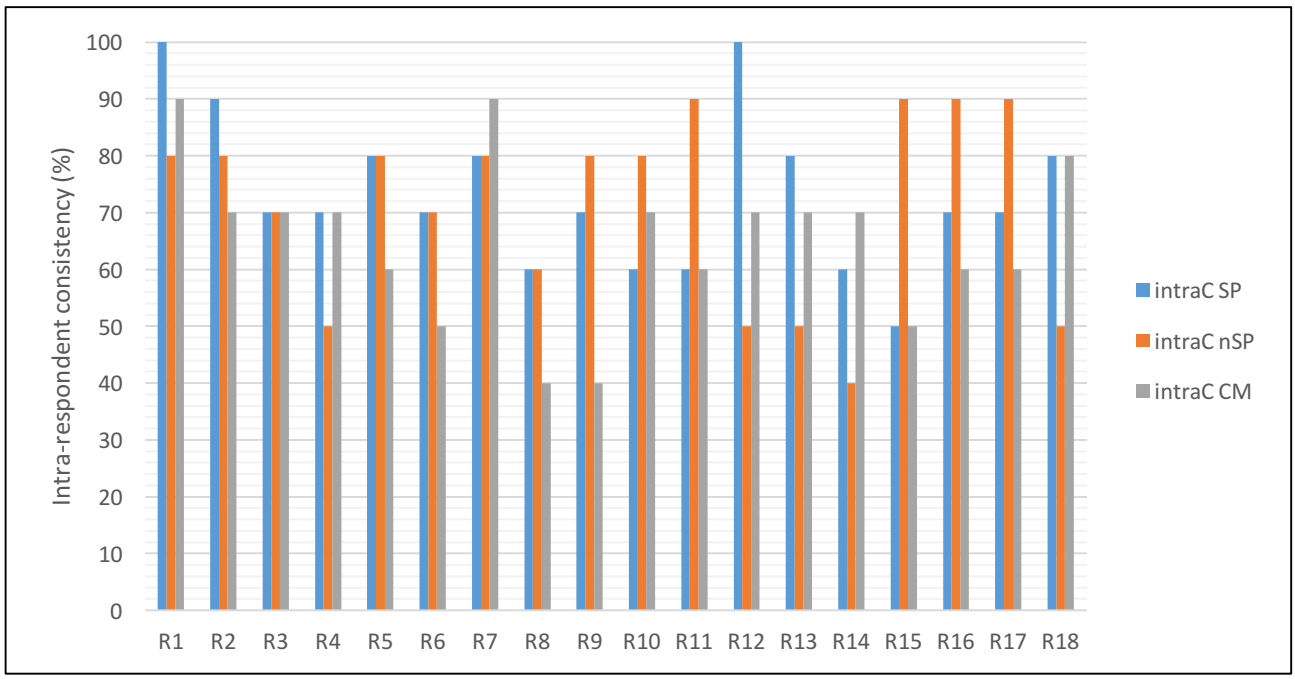
APPENDIX C

Mean success rates for different scales of modifications in ascending order (from 34 to 94 milliseconds) for items in speech block (SP), non-speech block (nSP) and cross-modal block (CM).



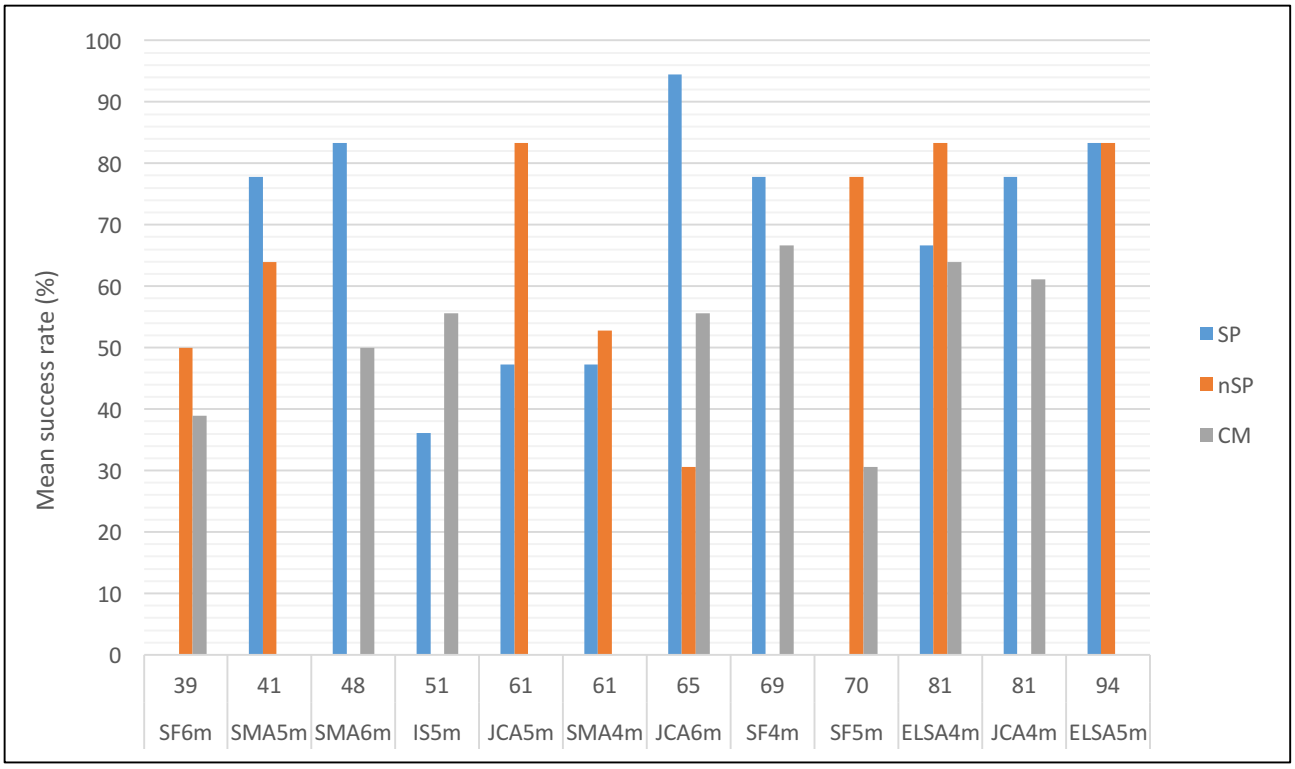
APPENDIX D

The calculations of intra-respondent consistency for the speech block (intraC SP), non-speech block (intra nSP), and cross-modal block (intra CM) for individual respondents (R1 – R18).



APPENDIX E

Mean success rates for recurring manipulated stimuli in speech block (SP), non-speech block (nSP) and cross-modal block (CM) with scale of modification (in milliseconds).



APPENDIX F

The answer sheet used in the experiment. The first, smaller chart served for the introductory training.

TRAINING		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠

TRAINING		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠

TRAINING		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠

SPEECH		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠
5	=	≠
6	=	≠
7	=	≠
8	=	≠
9	=	≠
10	=	≠
11	=	≠
12	=	≠
13	=	≠
14	=	≠
15	=	≠
16	=	≠
17	=	≠
18	=	≠
19	=	≠
20	=	≠
21	=	≠
22	=	≠
23	=	≠
24	=	≠
25	=	≠
26	=	≠
27	=	≠
28	=	≠
29	=	≠
30	=	≠

NON-SPEECH		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠
5	=	≠
6	=	≠
7	=	≠
8	=	≠
9	=	≠
10	=	≠
11	=	≠
12	=	≠
13	=	≠
14	=	≠
15	=	≠
16	=	≠
17	=	≠
18	=	≠
19	=	≠
20	=	≠
21	=	≠
22	=	≠
23	=	≠
24	=	≠
25	=	≠
26	=	≠
27	=	≠
28	=	≠
29	=	≠
30	=	≠

CROSS-MODAL		
item	same	different
1	=	≠
2	=	≠
3	=	≠
4	=	≠
5	=	≠
6	=	≠
7	=	≠
8	=	≠
9	=	≠
10	=	≠
11	=	≠
12	=	≠
13	=	≠
14	=	≠
15	=	≠
16	=	≠
17	=	≠
18	=	≠
19	=	≠
20	=	≠
21	=	≠
22	=	≠
23	=	≠
24	=	≠
25	=	≠
26	=	≠
27	=	≠
28	=	≠
29	=	≠
30	=	≠

Zhrnutie

Pozorovania z mojej didaktickej praxe naznačujú, že prozódia reči je jednou z posledných prekážok k dokonalému osvojeniu cudzieho jazyka. Moja učiteľská prax v oblasti anglického jazyka mi ukázala, že fonetické aspekty cudzieho jazyka sú v didaktickom prístupe často zanedbávané, a ich opomínanie prináša neželané efekty, keď napríklad študenti s vysokou úrovňou zvládnutia gramatiky alebo slovnej zásoby produkujú reč, ktorá svojim arytmickým charakterom len zďaleka pripomína prirodzený tok jazyka. Určite sa každému lektorovi už stalo, že nedokázal porozumieť svojmu študentovi angličtiny práve preto, že zo svojho jazykového prejavu vynechal prirodzenú prozódiu jazyka a striedanie prízvučných a neprízvučných slabík. Tieto aspekty by nemali byť zanedbávanou zložkou výuky jazyka, nakoľko sa vo významnej miere – tak ako dostatočné zvládnutie gramatickej alebo lexikálnej stránky – zúčastňujú na úspešnosti zvládnutia jazykového prejavu v jeho komplexnosti. Zároveň ako aktívny hudobník vidím dôležitosť detailného vnímania a náležitej produkcie rytmu aj v nerečovej oblasti. Keďže všeobecne je rytmus vytváraný striedaním prízvučných a neprízvučných elementov tak v reči ako aj v hudbe, prepojenie jeho hudobnej a rečovej formy rytmu v tejto diplomovej práci sa ponúka samo.

Teoretická časť práce sa zaoberá rytmom ako takým a jeho realizáciou v rečovom a hudobnom svete. Dôležitosť rytmu v umení rečníctva vyzdvihoval už Cicero. Filozofovia a antropológovia Darwin, Rousseau a Spencer vypracovali teórie o spoločný koreňoch jazyka a hudby. Niektoré hypotézy dokonca predpokladali, že prvotné jazyky boli spievané a používané predovšetkým za reprodukčnými účelmi. Hudba a jazyk sú v blízkom vzťahu a v rôznej miere podieľajú na prenose emócii a informácii, spoločne zdieľajú základné stavebné prvky ako rytmus a melódiu. Rytmus ovláda temporálnu stránku rečového aj hudobného prejavu, a melódia určuje tak intonačný pohyb reči ako aj hudobného motívu. Povel & Essens (1985) ukázali, že rytmicky organizované činnosti a udalosti sú jednoduchšie pre vnímanie a reprodukciu než tie, ktorým táto vlastnosť chýba. Je známe, že taktiež melodická stránka reči je dôležitou pomôckou pre sústredenie. Reč bez náležitého použitia intonácie pôsobí na človeka mdlo a hrozí väčšie riziko straty koncentrácie.

Koncept rytmu ako pravidelne sa opakujúcich rečových jednotiek bol založený na koncepte izochrónie. Fonetický výskum z polovice 20. storočia (Lloyd James, 1940; Pike, 1945; Abercrombie; 1967) kategorizoval jazyky sveta na dve kategórie podľa toho, či pravidelne

opakujúcim elementom v reči bola slabika alebo prízvuk. Túto impresionistickú hypotézu kategorizácie jazykov na základe ich rytmu sa následne snažili podoprieť viaceré fonetické výskumy (napr. Ramus, et al. 1999; Grabe & Low, 2002; Low, Grabe, & Nolan, 2000) pomocou takzvaných *rhythm metrics*, výpočtov predovšetkým vokalickej a konsonantickej variability, ktoré mali tento jav zachytiť. Výsledky týchto štúdii neboli natoľko univerzálne, aby presvedčivo interpretovali hypotézu rytmickej kategorizácie jazykov. Kritika na adresu *rhythm metrics* preto spočívala predovšetkým v pochybnostiach, či tieto kalkulácie nezachycujú v skutočnosti vlastnosti reči iné ako rytmus samotný (Kohler, 2009). Na základe toho Kohler navrhol presunúť zameranie výskumu rečového výskumu do roviny percepčných štúdii.

Koncept P-centier sa rozvinul v sedemdesiatych rokoch 20. storočia (Morton et al., 1976; Fowler, 1979) a jeho akustická podoba je stále predmetom výskumu v oblasti psychológie a fonetiky. Teória P-centier predpokladá existenciu takzvaných percepčných centier, ktorých výskyt predstavuje relevantné momenty pre poslucháčovu percepciu rečového rytmu. Podľa niektorých štúdii bola jeho pozícia v slove považovaná za danú na začiatku prízvučnej samohlásky v slove (Morton et al., 1976), a podľa iných (Rapp, 1971; Allen, 1972) zas jeho pozícia bola posunutá smerom k prevokalickým spoluhláskam.

Štúdie percepčných centier využívajú dve metódy – percepčnú a produkčnú. V percepčných výskumoch účastníci experimentu prispôsobujú trvanie intervalu medzi základným a cieľovým stimulom za účelom vytvorenia percepčnej izochrónie. V produkčných štúdiach zas účastníci synchronizujú svoju reč s pravidelne sa vyskytujúcim akustickým alebo vizuálnym signálom. Tieto štúdie poukázali na to, že umiestnenie percepčných centier môže byť podmienené okolitým prostredím, predovšetkým zložením konsonantickej skupiny, ktorá predchádza prízvučnú samohlásku, alebo dokonca aj samotnou dĺžkou tejto samohlásky. Kritika percepčných centier pozostáva predovšetkým na komplexnosť jazykového signálu a jeho potenciálne ohrozenie spoľahlivej interpretácie rečového rytmu a následnej identifikácie percepčných centier (Benadon, 2003).

Vývin metodológie resyntézy reči bol založený na predpoklade, že vnímanie lingvistického rytmu je založený na schopnostiach, ktoré nie sú priamo závislé na jazyku samotnom. Resyntéza mení tok reči na nerečový signál, odstraňuje lexikálnu a fonotaktickú informáciu, a zachováva iba niektoré vybrané atribúty reči ako napríklad rytmus alebo intonačnú krivku.

Počiatočné percepčné štúdie s resyntetizovanou rečou boli uskutočnené na novorodencoch. Častou metódou bolo použitie nízkopásmového filtra na odstránenie segmentálnych informácii (Mehler et al., 1988; Nazzi et al., 1998), avšak táto metóda bola kritizovaná ako nedostatočná, pretože neumožňuje precízne oddeliť jednotlivé zložky reči (Ramus, et al., 1999). Následné experimenty boli postavené na dôslednej delexikalizácii rečového signálu pomocou nahradenia jednotlivých hlások za univerzálne segmenty. Týmto spôsobom sa podarilo zachovať rytmus a intonáciu, zatiaľ čo ostatné, nežiadané rečové informácie boli spoľahlivo odstránené. Výsledky experimentu poukazujú na zvýšenú úspešnosť v rozlíšení jazykov podľa ich rytmických vlastností.

Viaceré štúdie poukazujú na spoločné kognitívne procesy potrené k spracovaniu hudby a reči (Patel, 1998, 2003a, 2008; Lerdahl and Jackendoff, 1983; Jackendoff, 1989). Tie spadajú do oblasti syntaxe, melodickej a rytmickej organizácie toku hudby a reči. Bolo taktiež poukázané na fakt, že hudobný tréning zvyšuje citlivosť na javy ako melódia a rytmus, ktoré sa tiež prejavujú v reči v jej intonačnej a rytmickej rovine (Tzounopoulos & Kraus, 2009). Na druhej strane bol preukázaný vplyv lingvistického tréningu v podobe výuky cudzích jazykov s odlišnými rytmickými vlastnosťami na spracovanie rytmu celkovo (Roncaglia-Denissen, et al., 2016).

Cieľom tejto práce bolo empiricky preskúmať koncept rytmickej senzitivity a jeho aplikáciu na krátke rečové a nerečové frázy. Zároveň bol vplyv hudobného tréningu, súčasného alebo predchádzajúceho, na vnímanie rytmických odchýlok jedným zo sekundárnych zameraní práce. Na základe poznatkov z predchádzajúceho výskumu percepcie rytmu v oblasti reči a hudby sme sformulovali niekoľko hypotéz, ktoré sa týkali úspešnosti identifikácie rytmických odchýlok v percepčnom experimente. Predpokladajú, že účastníci s absolvovaným hudobným tréningom budú v teste úspešnejší, že manipulácie rytmu v prízvučných slabikách budú percepčne výraznejšie než tie v neprízvučných, že manipulácie rytmu budú evidentnejšie v rečovom než v nerečovom signáli, a že pozícia rytmickej manipulácie ovplyvní výsledky experimentu – čím neskôr vo fráze sa manipulácia vyskytne, tým nižšia úspešnosť jej identifikácie. Naše výskumné otázky sa venovali najmenšej možnej postrehnuteľnej rytmickej odchýlke v rečových a nerečových frázach, celkovej dĺžke fráze, a potenciálnym faktorom, ktoré môžu účastníkom v experimente zabrániť v úspešnej identifikácii rytmickej modifikácie.

Za účelom otestovania hypotéz a zodpovedania výskumných otázok bol vytvorený nasledovný experiment. Na preskúmanie rozdielov v senzitivite voči rečovému a nerečovému rytmu bolo

potrebné nájsť vhodné nahrávky a vytvoriť ich nerečové ekvivalenty. Tri krátke frázy piatich rôznych hovorcov s dĺžkou 4-6 slabík boli vybrané z dlhších nahrávok z Pražského fonetického korpusu. Ich nerečové ekvivalenty boli vytvorené pomocou dvoch krátkych zvukov woodblocku určených pre reprezentáciu prízvučnej a neprízvučnej slabiky. Pomocou softvéru na úpravu zvuku boli perkusívne signály zosynchronizované s predpokladanou lokalitou P-centra (Cummins & Port, 1998; Fowler, 1983; Scott, 1993; Scott, 1998), a to v mieste najväčšieho zlomu amplitúdovej obálky samohlásky v prízvučnej slabike.

Manipulácie rytmu v rečových aj nerečových frázach boli prevedené algoritmom PSOLA, ktorý digitálnou technikou *overlap-add* rozdeľuje zvukový signál na malé prekrývajúce sa segmenty. Tie následne replikuje a kombinuje, aby zmenil trvanie signálu. Týmto spôsobom sme predĺžili trvanie slabík o 50%. Plynulosť rečového prejavu bola kontrolovaná a tie manipulácie, ktoré by ju narušovali, boli vyradené z experimentu.

Experiment pozostával z troch rôznych blokov, ktoré boli postupne predkladané respondentom v percepčnom experimente: rečovom, nerečovom a cross-modálnom. Každý blok obsahoval tridsať párov fráz. Prvá fráza bola referenčná, a druhá bola buď identická alebo obsahovala rytmickú manipuláciu. Prvých dvadsať fráz v každom bloku bolo unikátnych a zvyšných desať bolo kontrolných prvkov, ktoré slúžili na následný výpočet konzistencie vo výsledkoch pre daného respondenta. Rečový blok obsahoval iba rečové fráze, a nerečový blok obsahoval iba nerečové, t.j. perkusívne fráze. Páry v cross-modálnom bloku boli zložené s rečovej frázy a jej (rytmicky ekvivalentného alebo rytmicky modifikovaného) nerečového ekvivalentu.

Výsledky experimentu poukázali na niektoré rozdiely vo vnímaní rečového a nerečového rytmu. Konzistencia výsledkov ako aj najvyššia úspešnosť v identifikácii rytmických odchýlok bola zaznamenaná pre rečový blok. Najlepšie výsledky sa v rečovom bloku prejavili, keď sa rytmické manipulácie nachádzali na prvej alebo tretej slabike frázy, a keď táto slabika bola prízvučná, alebo bola takouto slabikou priamo nasledovaná.

Najlepšie výsledky pre nerečové frázy boli naopak dokumentované, keď sa rytmické manipulácie nachádzali na druhom beate perkusívnej frázy, a keď tento beat bol neprízvučný, alebo bol takýmto beatom nasledovaný. Tieto výsledky priamo protirečia hypotéze, ktorá predpokladala, že identifikácia rytmických manipulácii bude uľahčená prízvučnosťou daného elementu.

Nadpriemerná identifikácia rytmických manipulácii sa vyskytovala, keď ich dĺžka prekonala hranicu 60-70 milisekúnd. Táto hranica bola o zhruba 5 milisekúnd nižšia pre nerečový blok než pre rečový blok.

Hypotéza, že cross-modálny blok bude predstavovať najťažšiu časť experimentu, sa potvrdila. Porovnávanie dvoch fráz rôzneho typu bolo náročné pre účastníkov, ktorí sami potvrdili neistotu v tom, či rečové ekvivalenty korešpondovali s ich čiste rytmickou reprezentáciou. Aj keď umiestnenie beatu v slabike bolo v súlade bežne prijímaným umiestnením s percepčných centier na začiatku samohlásky (Cummins & Port, 1998; Fowler, 1983; Scott, 1993; Scott, 1998), pomerne nízke výsledky v cross-modálnom bloku poukazujú na potenciálnu variabilitu percepčných centier vzhľadom na iné premenné, ako sú prevokalické elementy, trvanie samohlásky, celej slabiky, alebo tvar spektrálnej obálky (Fox & Lehiste, 1987; Scott, 1993; Harsin, 1997; Howell, 1988; Pompino-Marschall, 1989).

Vplyv na výsledky experimentu malo taktiež poradie blokov a dĺžka fráze. Podstatný pokles v úspešnosti nastal v prípade šesťslabičných fráz, čo naznačuje klesajúcu tendenciu ľudského ucha zapamätať si rytmické vzorce s narastajúcou dĺžkou fráze.

Hypotéza, ktorá sa týkala vplyvu hudobného tréningu na výsledky v experimente, sa nepotvrdila. Ani hudobný tréning, ani účasť v hudobnom kolektíve neovplyvnil výsledky. Práve naopak, často to boli účastníci s nulovou skúsenosťou s hudobným tréningom, ktorí dosahovali najvyššie výsledky v percepčnom experimente.

Jazyková úroveň angličtiny ovplyvnia výsledky iba v rečovom bloku. Čím vyššiu úroveň angličtiny účastníci mali, tým vyššiu úspešnosť dosahovali. Tieto výsledky poukazujú na fakt, že dlhodobá skúsenosť s jazykom prízvučného rytmu môže ovplyvniť schopnosť rozlíšiť rytmické odchýlky (Lidji et al., 2011). Predpokladá sa, že účastníci s lepšou úrovňou jazyka sú vystavení dlhší čas prirodzenému toku reči a dokážu rozoznať jej prirodzený rytmus citlivejšie než účastníci s nízkou úrovňou daného jazyka.

Táto diplomová práca používala dva druhy krátkych rytmických fráz a skúmala schopnosť hudobne trénovaných a netrénovaných účastníkov rozoznať rytmické manipulácie na rôznych miestach a rôznych dĺžok týchto fráz. Analýza konkrétnych segmentálnych podmienok a prostredia v rečových frázach bola mimo možností tejto štúdie. Nakoľko nie je vplyv iných nepreskúmaných faktorov na vnímanie rečového rytmu a umiestnenie percepčných centier vylúčený, ich výskum v nasledujúcich rokoch je očakávaný.