UNIVERSITY OF JYVÄSKYLÄ

Speech recall and word recognition depending on prosodic and musical cues as well as voice pitch

Anna Rozanovskaya
Taisia Sokolova
Master's Thesis
Music, Mind & Technology
University of Jyväskylä
August 2011

2011

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty	Laitos – Department
Faculty of Humanities	Music Department

Tekijä – Authors
Anna Rozanovskaya
Taisia Sokolova

Työn nimi – Title
Speech recall and word recognition depending on prosodic and musical cues as well as voice pitch

Oppiaine – Subject	Työn laji – Level
Music, Mind and Technology	Master
Aika – Month and year	Sivumäärä – Number of pages
September 2011	83

Tiivistelmä – Abstract

Within this study, speech perception in different conditions was examined. The aim of the research was to compare perception results based on stimuli mode (plain spoken, rhythmic spoken or rhythmic sung stimuli) and pitch (normal, lower and higher). In the study, an experiment was conducted on 44 participants who had been asked to listen to 9 recorded sentences in Russian language (unknown to them) and write them down using Latin letters. These 9 sentences were specially prepared using different phonetic environments, voice pitches, rhythmic structures and presentation modes (sung/spoken).

The analysis showed that recall ability (for both the amount of remembered and recall accuracy) had been affected by the mode of stimuli: sung stimuli had been remembered better than spoken ones, even compared to the rhythmical spoken condition. Segmentation accuracy turned out to be enhanced only with introduction of rhythm, without any significant improvement for the sung stimuli. Voice height was found to influence perception of the phonetic image, affecting recognition of vowels in stressed positions and consonants both in stressed and unstressed positions. This effect was especially strong for the higher pitch and plain spoken stimuli.

The data obtained in the study and experimental design developed for it may be used both for further research purposes and for creating educational settings.

Asiasanat – Keywords
music & language, prosodic cues, music cues, language through music, word recognition, speech segmentation, speech recall, text memorizing, voice height

Säilytyspaikka – Depository

Muita tietoja – Additional information

Table of contents

List of tables

List of graphs

1 Introduction

As representatives of two different worlds — the world of classical music, with its fundamental insight into musical laws, and the world of theoretical and applied linguistics, with its focus on language structures and phonetic rules, — the authors of the study have a rare opportunity of combining effort, knowledge, experience and passion for the purpose of developing, conducting and analyzing the results of an interdisciplinary research.

The scope of the authors' interest includes possibilities and perspectives of incorporating musical elements (like songs) into language-oriented educational settings. One of the relevant issues is the problem of segmentation and word recognition in the foreign speech, as well as eliminating the task of remembering foreign texts. Different ways of introducing songs into the language studying process are relatively widely discussed and used in practice. Some researchers provide clear evidence supporting the vision that songs can somehow contribute to the foreign language learning practices (for example, facilitate memorization of words), while others dispute this claim, arguing that presenting songs, as a form of authentic information, has no advantage comparing to other forms of input.

As shown later, there are quite a number of studies considering different positive aspects songs can have for the process of second language acquisition. These are, for example, gaining attention, introducing foreign language prosody, help in memorization of words or structures, involvement of foreign culture elements. However, research on how music, in the form of songs, influences the text segmentation ability is lacking. Nonetheless, this is a fundamental issue. It is clear that no proper intercultural communication can take place without accurate speech segmentation by the parties. The problem of recognizing sounds of foreign language stays acute even for advanced language learners, while beginners are often simply unable to identify the few words they already know in a fluent speech flow. On the other hand, conventional language teaching methods offer only a slightly varying set of audition techniques, which generally are limited to: (a) passive repeated exposure (listening) and (b) exposure (listening) along with simultaneous reading.

At this point, it is necessary to define terms "word recognition" and "speech/song segmentation", as they are used within this study. In our terms, "segmentation" refers to the ability of a human adult, with normal hearing and mentally healthy, to distinguish boundaries between such units like word or 'word + preposition' in a fluent speech stream or a song in a

foreign language. Then, "word recognition" means the ability of a human adult to extract certain words from a fluent speech/song stream. It is more likely to refer to a language, lexical, phonetic and rhythmical structure of which is somehow familiar to the subject. In this paper, however, a word is considered "recognized" if its boundaries and phonetic image are reproduced accurately by the listener. By "phonetic image" we understand phoneme combination perceived as a unit.

It is well known that melody added to a text changes its prosodic properties. The nature and valence of these changes are not clear and seem to be strongly dependent on a) the melody's properties and b) the target language characteristics. It is not the aim of the current study to look at how different melodic contours and rhythmical patterns affect lyrics segmentation in a particular language. Nevertheless, we are here comparing segmentation outputs for plain text stimuli with the outputs for stimuli sung to an "average" melody. By "average", we understand a relatively simple but not primitive melody, which sounds natural to speakers of a particular language.

To check the assumption about purely prosodic cues (i.e. that language prosody cues do influence speech segmentation while music cues do not), supported by some studies referred to later, it was decided to introduce a rhythmically organized piece of text without musical features. Using stimuli in the form of rhythmic text, which exactly resembled the sung samples, with the exception of melody (which was absent), allowed to test the difference in plain text vs. sung stimuli perception and be sure that it was ecologically valid.

Another dimension of the research dealt with the problem of segmenting input stimuli at different pitch levels/tones. There are some studies demonstrating that considerable pitch shift of the signal may lead to a changed perception of sounds. While the general view of perception mechanisms claims, that the human perception system tends to cut off extraneous features from an acoustic signal and focus on its invariant features (see the Theoretical considerations chapter), it is unlikely that an extreme shift in pitch may be ignored and leave the signal perception unchanged.

Although the field of pitch perception, in general, has been investigated quite thoroughly, unfortunately, there is a lack of studies examining the effect that pitch/tone of an utterance/song may have for the segmentation and word recognition ability. By "pitch", we mean, here, a range of frequencies in correspondence to piano registers. Thus, "high voice" corresponds to the lower level of the upper register of the piano keyboard; "middle voice" relates to the middle register of the piano keyboard; and, finally, "low voice" refers to the

upper region of the low register of the piano keyboard. To develop a working research mechanism allowing analyzing this aspect of perception, we used a single ordinary female voice recording aligned in pitch with the required frequency range (pitch) using modern sound processing tools.

Besides these two aspects, speech segmentation (finding word boundaries) and phoneme perception, connected with the conception of word recognition, the other relevant ability involved into the learning process is remembering, or “ability to recall”. Memorizing words, language constructions and, wider, texts in foreign language is essential for learning progress, and it is very important to find ways to facilitate the task, make it more interesting, easy and effective. While research demonstrates support for various opinions on the capacity to memorize songs as opposed to spoken texts, it is particularly intriguing to put the sung stimuli into the same context as spoken rhythmic stimuli using prosodic cues and find out, whether such musical features as melody can gain additional benefits for the text recall ability.

This research is intended to observe dependency of the human ability to recall and segment speech and recognize words on various controlled conditions. These conditions include type of input ("mode") – spoken plain text, sung rhythmic text and spoken rhythmic text, – and voice height. The aim of the study is to reveal such possible correlations, should they take place.

The Chapter 2 of the paper presents detailed theoretical considerations on the matters discussed above, as they are seen from musicological and linguistic perspectives. Quite a number of various studies demonstrating different perspectives are described there. In the Chapter 3, we shortly introduce the research questions and working hypotheses of the study.

The empirical approach is pictured in the Chapter 4, the first section of which introduces in detail the research method, variables to be observed, experiment description (including linguistic and musical approaches to the stimuli creation, experiment design and procedure), information on participants and statistical methods used in the analysis. The second section of the chapter contains analysis of the output data. It includes description of variables calculation methods, correlation tables and results of various tests, as well as some conclusions on the analysis. It also describes qualitative data extracted from the questionnaires filled in by the participants. This questionnaire was designed in order to give

the personal information reflecting participants' opinions on the experimental setting and different types of stimuli.

Expected challenges were connected with the complexity of the task for participants, as they were asked to write down heard stimuli, which some might find stressful and uncomfortable. Another difficulty was about analysis, namely, converting the raw data into countable variables.

The setting of the research was quite complex and integrated multiple planes. We wanted to analyze various aspects of recalling heard foreign texts and segmentation/word recognition. This job required much effort, time and creativity, as we had to produce stimuli "from the ground up" and without someone's extraneous help.

During the work on this thesis the labor was divided between us two as follows: Taisia was analyzing the literature, dealing with music aspects of the language perception, and wrote the Parts 2.2 – 2.6 of the theoretical section; while Anna prepared the linguistic review of the topic (Part 2.1). Apart from this, Anna also prepared descriptions for the Section 4 (except for the Part 4.1.2.3). The experiment was conducted by our united effort where Anna prepared the linguistic (word) content of the stimuli, while Taisia reasoned out their music aspect, and both were equally involved into the process of obtaining, entering and analyzing the data.

2 Theoretical considerations

2.1 Linguistic aspects of speech recognition

2.1.1 History of the word recognition research in a nutshell

The research on speech perception and word segmentation embraces a wide range of studies conducted over the past half a century. Since 1950s, evident progress has been achieved in this field: from early studies on categorization and distinction of speech sounds, to works analyzing processes underlying spoken word recognition and acquisition the ability to perceive native language by humans. A selective review "Speech Perception and Spoken word Recognition: Past and Present" by Peter Jusczyk and Paul Luce (Jusczyk & Luce, 2002) gives a good overview of what has been done in this area. Thanks to that comprehensive review, there is an opportunity to reduce the exhaustive and time-consuming stage of searching for retrospective research and address directly only to the most remarkable works in this field.

At early times of speech perception research, the general adopted idea was to see the language as a hierarchical structure with a number of distinctive levels. Consequently, to obtain an accurate description of the structure, a scrupulous description of each of these levels was required. Basically, in those terms, a description of the language would include phonetic level (acoustic properties of utterances mapping onto phonetic segments), phonemic level (phonetic segments mapping to particular phonemes), morphemic level (combinations of phonemes forming morphemes) and syntactic level (combinations of morphemes forming sentences). Most of the studies at that time focused on three issues critical for understanding processes of transforming acoustic signal into phonetic segments: invariance, constancy and perceptual units.

In two words, the concept of **invariance** suggests that in the language, as in a system with a certain structure, each phonetic segment might be determined by acoustic properties of a unique set. However, situation in a natural language is much more complex. Research (Delattre, Liberman & Cooper, 1955) showed that there were no obvious shared acoustic properties that could specify the same consonant (in particular) in different contexts. Delattre et al' s study of perception of the consonant [d] in different realizations made it clear that

acoustic features of the consonant were highly influenced by the following vowel. This effect is due to the phenomena of coarticulation (simultaneous production of the consonant with the following vowel). This and some other works (Liberman, DeLattre & Cooper, 1952) with their findings made doubtful the mere idea of discriminating phonetic segments based on the invariant acoustic properties.

Then, speaking about the issue of **constancy**, the variable realization of phonetic segments depending on their phonetic context is not the only variability. Speech perceptual system faces also differences in the acoustic properties of sounds produced by different talkers intending to say the same sound. These may be pitch differences in male and female voices, and other differences due to varying sizes and shapes of talkers' vocal tracts. It may happen that production of a word by a particular talker resembles the production of a different word by another talker, while sounds different from the production of the same word by the second talker (Ladefoged & Broadbent, 1957). Nevertheless, two productions of the same word by the same talker may differ significantly in pitch and other acoustic properties (Fernald, Taeschner, Dunn, Papousek, Boysson-Bardies & Fukui, 1989; Kuhl et al., 1997). Still, human recognition system seems to cope easily with the challenges (Creelman, 1957; Verbrugge, Strange, Shankweiler & Edman, 1976) and recognize the same words irrespective of talkers producing them.

There was a tendency in the early studies to consider the elementary **perception unit** as a match to the phonetic segment, the smallest unit capable to distinguish two different forms of a word, based on minimal contrasts in certain features to other units (Jakobson, Fant & Halle, 1952). With the introduction of the pattern playback synthesizer using data from spectrographic analysis of speech for speech sounds generation, researchers were able to look separately at different elements of acoustic signal and try to understand their effect on speech perception (Cooper, Delattre, Liberman, Borst & Gerstman, 1952; Cooper, Liberman & Borst, 1951).

At this point, a very important notion for speech processing should be introduced: *formant*. Formants are the name for bands of energy concentrated at different acoustic frequencies that can be seen on spectrograms. Formants correspond to the natural resonant frequencies which the vocal track produces in speech. The phenomena of coarticulation revealed the effect when phonetic segments were not necessarily assigned to different particles of the acoustic signal. This stimulated many researchers to try and identify the elementary unit of perception.

By now, contradictory data has been acquired. Thus, some researchers (Savin & Bever, 1970; Massaro, 1972) found that detecting syllables took less time for listeners than detecting phonemes. Consequently, they suggested syllable to be the minimal unit of perception. However, this view was contradicted by other research showing that in different conditions faster detection times were observed for phonemes (Cutler, Norris & Williams, 1987; Healy & Cutting, 1976; Mills, 1980; Swinney & Prather, 1980) or units bigger than syllables (McNeill & Lindig, 1973). At this stage, there is no agreement as to what should be considered the elementary perception unit.

Eventually, in their search for the correlates between minimal language units and perception units, researchers' attention was gained by the issue of phoneme perception. Given that phonemes were elementary phonetic elements used for building words, an assumption was made, that recovery of phoneme sequence in the word (which we call *phonetic image* in terms of this paper) was crucial for perceiving words.

It was found that so called *categorical perception* was used for speech recognition. This, in particular, meant that for listeners it was easier to make the distinction between sounds from different sides of a phoneme boundary, than between sounds lying within the same phoneme category (Liberman, Harris, Hoffman & Griffith, 1957, p. 358). This claim was surprising, because, for other types of acoustic signal, listeners usually had revealed the ability to make much finer distinctions. For example, listeners were found to discriminate about 1200 pitch differences in the frequency range of 100 to 8000 Hz, having only around seven labels to name them (Pollack, 1952).

Thus, categorical perception deals with the phenomena of sharp distinction between phonetic categories and poor discrimination within these categories. However, categorical perception is not limited to this dimension and involves also such phonetic contrasts as voicing (Liberman, Harris, Kinney & Lane, 1961) and manner of articulation (Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975).

There is a widely shared view that categorical perception is unique to speech. This view is supported by the findings that while changes in speech sounds are perceived categorically, the non-speech contrasts with similar acoustic changes are perceived continuously (Liberman, Harris, Eimas, Lisker & Bastian, 1961). Nonetheless, exploration of the complex non-speech stimuli perception has shown an evidence of categorical perception also in that domain.

The discussion about minimal perception units and, generally, about word recognition mechanisms is essential for the purposes of this study, as this is through it that we look at perceptual constancy of the same units of a foreign language, produced in different pitches and in different speech modes. For this purpose, we have to define units to be measured, considering their phonetic environment.

2.1.2 Prosodic cues for speech perception

As long as our study aims to distinguish music and prosodic cues for speech perception, it is necessary to give a brief look at research in the field of language prosody. There have been quite a number of studies examining the influence, which rhythmic structure and metrical expectations, etc., have on speech perception.

It has been shown by many researchers that in certain conditions listeners tend to use one or another segmentation cue. However, none of these cues may be considered absolutely reliable, while some studies show that listeners usually rely on a combination of cues for speech processing. For example, Sanders and Neville (2000) investigated the ability of young listeners to use different lexical, syntactic and stress-pattern cues for speech segmentation. Results revealed that participants used multiple cues at a time, and could do it quite flexibly (Sanders & Neville, 2000).

Then, Quen'e and Port (2005) found that rhythmic regularity could enhance speech perception (word recognition, to be more precise). Studying expectancy for rhythm and meter in participants, they revealed that clear rhythmic regularity had a strong influence on word perception: clearer rhythmical structure led to better spoken word perception. On the other hand, it was found that metrical regularity had no or little effect on speech perception. Authors claimed that the main finding of their research was the idea of "attentional" rhythm used by listeners for speech processing. Their study also revealed that clear rhythm facilitated speech communication. (Quen'e & Port, 2005).

Another study, by Zheng & Pierrehumbert (2010), analyzed the influence of prosodic expectations on speech perception and used for this purpose different (dactylic, iambic, and trochaic) sentences at slow and fast presentation rate. It was found that prosodic cues (lengthening) had a bigger effect for strong syllables in all experimental conditions. It might lead to a suggestion that strong syllables provided some perceptual advantages for recognition and identification processes. In this work it was also revealed that metrical expectations also

played some positive role for the identification task that led to better performance, allowing participants to focus their attention on metrically prominent syllables.

Nevertheless, the important finding of that study was that stress was the most crucial component in the process of speech perception. Another influence that was found, was that of the meter (less significant, though). The central point was that people demonstrated better prosodic cues detection for strong syllables. It meant that stressed words generally provided the most important semantic information for speech perception. Stress in the speech awoke more attention that helped to comprehend sentences. This study also showed that clear rhythmic structure provided a good framework for speech perception and facilitated the process of retrieving information from the speech stream (Zheng & Pierrehumert, 2010).

Other cues that may also have a positive influence for speech segmentation are transitional probabilities (some kind of likelihood that one element will follow another) between speech units. They provide enough information to discriminate the word boundaries, at least on the first stage of language acquisition. For example, Aslin, Saffran, & Newport (1998) in their work "Computation of conditional probability statistics by 8-month-old infants" found that newborns could segment continuous stream of words without using any acoustic or prosodic cues. Further application of language-specific prosodic cues also facilitated the process of finding word boundaries. Saffran, Newport & Aslin (1996) in their study "Word segmentation: The role of distributional cues" investigated the role of distributional cues for speech perception. Results showed that people were able to remember and segment words in conditions where only transitional probabilities were available as perceptional cues. Results also revealed significant progress in performance when certain prosodic cues were added, which demonstrated a critical role of prosodic cues for speech processing.

2.2 Using music as a tool in educational settings – pro and contra

2.2.1 Pros

First question that has to be discussed before considering the whole theme more in detail is: can music make the language-learning process more effective? Are there any musical features that might contribute to this process? Are there any reasons for using it in educational settings? In this connection, it seems important to give an overall representation

of possible ways, in which music may be used in foreign language studies. There are some general considerations about what role music in the form of songs can play in the language learning process: gain attention and awareness, make stresses and accents, evoke interest and curiosity, intrigue, facilitate memorizing things, and help with pronunciation. It also helps with understanding the vocabulary out of the context; allows to see the beauty and variety of a foreign language; allows to gain more interest for a foreign language; allows to see success after many repetitions of songs; improves the solidarity feeling of a class; offers the opportunity to learn more about the country of the language and it's culture; furthers the acoustic learning and helps with grammar.

The process of second language acquisition is quite complex. It includes such components as attention (focusing on the subject) –> initial learning (using short-term memory) –> training (getting familiar with the word / grammatical construction) –> recalling (extracting information from the memory using long-term memory) –> and integrating (ability to freely operate the second language units).

According to Ashcraft (2006), strong mnemonic device demonstrates three main principles: it gives a particular structure to incorporate information into the existing memory framework. Then it helps to incorporate this information. And finally, the mnemonic device facilitates retrieval of information (Ashcraft, 2006).

From the very first sight, it seems that music (especially in the form of songs) can play a role in some of the above mentioned components, so it may be used as an effective mnemonic device. Generally, most studies use songs or melodies – singing for the purpose of learning vocabulary, grammar or remembering a piece of text. This, we may call "active participating" in the learning-through-music process.

Introducing music (especially singing) into the learning process may provide an additional motivation to learn. Wigram and Gold (2006) observed that even children with communication difficulties enjoyed musical activities and often felt themselves more comfortable and less isolated – due to music. Their study stressed the benefits of music for facilitating social responsiveness, levels of engagement and verbal responding. It showed, as well, music's assistance as a relaxing and motivating tool (Wigram & Gold, 2006).

There are quite a number of studies describing other effects music has on the second language learning process. For example, it appears that hearing a melody of a well-known song can cue the text and, vice versa, hearing the text can cue the melody. In her work, Wallace (1994) asked: "Why are the text and melody effective cues for each other even after

long retention intervals?" As possible answers, she described several hypotheses that could explain this phenomenon. Firstly, she suggested that combining text with a melody could add more uniqueness or connections and make them easier accessible. An alternative explanation could be, that repetition of songs itself established a strong, stable memory. And finally, Wallace assumed that easier retrieval of songs might be explained by the theory, that songs and texts were stored or processed neurologically in different ways (Wallace, 1994).

Another study examining the effectiveness of melodic-rhythmic mnemonics, as an aid for short-term memory, was undertaken by Gfeller (1983). It used variables of group membership (learning disabled and normal students) and rehearsal mode (musical and verbal) with 30 normal and 30 learning disabled boys, in the age between nine and twelve years. Results showed that extended rehearsal of the musical mode, combined with strategy's modeling and cuing, had provided significantly greater recall for both normal and disabled students. According to Gfeller, this study indicated that musical mnemonics might serve as a useful tool for retention for both learning disabled and normal students (Gfeller, 1983).

Rainey & Larsen in their work "The Effect of Familiar Melodies on Initial Learning and Long-term Memory for Unconnected Text" tested how music (in a form of familiar melodies) could serve as an effective mnemonic device. The experiment showed, that participants exposed to the sung version of data required fewer trials to relearn this data a week later than did participants who had listened to the spoken version. This study once again argued that music might have a positive impact for relearning information, and could be used as a source of data and techniques (Rainey & Larsen, 2002).

In another study, McElhinney and Annett (1996) proved the influence of music on recall of verbal information using unfamiliar tunes and lyrics. Results showed that using music (songs) to help recall had been very effective. Participants had better overall recall when a song was used to present information. Tests showed that the amount of words per unit recalled by song group was significantly higher than that of the spoken group (McElhinney & Annett, 1996).

An experiment conducted by Chazin and Neuschatz (1990), however, showed that information did not have to be familiar. They tested the effect of music as a mnemonic instrument for recall of unfamiliar scientific information among 8-year-olds children and young adults. Results revealed that there was higher recall of information with the musical condition than with the traditional lecture (Chazin & Neuschatz, 1990).

Wallace (1994) also acknowledged that sometimes song's melody could make text perception easier. Wallace's experiments indicated that the role of melody was not limited to just rhythmical information. Rich structure of music linked words with phrases, identified the length of lines, defined stress patterns and added emphasis. Additionally, it helped the listener to focus on surface characteristics. The general conclusions by Wallace (1994) are the following: material is affected by presence of structural characteristics in this material, by simplicity of finding and perception of those characteristics, and by the contribution those characteristics make for organizing, constraining or cueing. Music seems to accomplish all three of these conditions.

Salcedo (2010) examined effects of using songs in three key areas of the foreign language learning process: students' ability to recall song lyrics, delayed recall of song words, and, finally, occurrence of involuntary mental rehearsal din (musical din) for sung and spoken text. Results of the study showed, that scores of students in the music group were much higher than those of students in the text group. The second result concerning the delayed text recall was negative. The findings revealed no difference in delayed text recall. But the last question about the occurrence of involuntary mental rehearsal (din) showed a significant advantage of the sung material: participants from the musical treatment class reported higher occurrence of din. It was obvious that music had an advantage in increasing din occurrence (Salcedo, 2010).

Krashen (1983) formulated that din might be a sign that language acquisition occurs (Krashen & Terrell, 1983). So, it's possible to suggest that increasing of din by using songs may have a positive influence on second language acquisition. In any way, this aspect is worth to be investigated more thoroughly.

Nevertheless, findings, provided by Salcedo (2010), indicated that music had a clear pedagogical value, showing that usage of music and songs for language acquisition appeared to be a more efficient way to activate mental rehearsal that led to more successful stimulation of the language learning process (Salcedo, 2010)

Butzlaff (2000) performed a meta-analysis of studies from 1950 up to and including 1998 which investigated the effect of music on developing reading abilities in children. The meta-analysis of 6 experimental and 24 correlational studies revealed, that 80% of the studies had shown some positive influence of music, as well as demonstrated correlation between the reading ability and music instruction, which became most clear in correlational studies (Butzlaff, 2000). Butzlaff's study supported the use of music in reading instruction. However, the small number of experimental studies can be seen as a limitation of this meta-analysis.

2.2.2 Cons

Despite the long-held belief that music may facilitate the learning process, and despite crucial number of studies claiming that music must serve as an effective tool in educational settings, there are also works providing evidence against it. One of the main questions may be formulated as "why and how music could help with, for example, word recall, when there is additional information (e.g. melody) to remember in the song comparing to the plain text?" That must require from the participant to perform a double task, providing additional difficulties to the learning process. In our case, it could be quite profitable to consider also studies showing negative results of using music, and to analyze, in which cases these negative results took place or (when speaking about songs) which properties of a particular song might prevent better remembering and recalling process.

For instance, study fulfilled by Kouri and Telander (2008) didn't prove the suggestion that readings of a sung story book would improve the abilities of story understanding and narrative retelling in children with speech and language delay. Thirty pre-school and first grade children with speech and language delay were exposed to book readings, in either sung or spoken condition, and then asked to repeat stories and answer story understanding questions. The results didn't show any significant difference in story retelling and comprehension competencies between both conditions. Authors explained that for children with speech and language delay sung condition might have introduced an additional amount of information to process, and as a result music had drawn their attention away from the semantic component of the sung story. On the other side, it was found that sung condition had enhanced the participants' story narratives in terms of vocabulary usage (Kouri & Telander, 2008).

This finding referred to another research, conducted by Kouri and Winn (2006), examining how singing affects quick incidental learning of new vocabulary terms. Though outcomes demonstrated no significant difference in target lexical items comprehension, clear positive effects of music on vocabulary learning were revealed (Kouri & Winn, 2006). This is supported also by another study, by Salcedo (2010), mentioned above, in which three different songs were used. Results of the study showed a significant difference between "music" and "prose" groups, in favor of the music condition, but only for the songs 1 and 3. The outcomes for the song 2 did not show a significant difference between the compared groups. Author suggested that this might be due to the song chosen. It was a romantic ballad

dating back to 1951, with complicated and unusual vocabulary, accompanied by an orchestra, which might be quite good for enjoyment, but too complex to remember.

The result corresponding to the delayed text recall was also negative. The findings revealed no difference in delayed text recall. However, Salcedo (2010), appealing to Bygrave (1995), claimed that the reason might be about too short term between the experiment and the delayed recall task, and that with a longer period the test might demonstrate a more significant music effect.

Presentation rate may also influence the process of memorizing text when it is set to a melody (Kilgour;Jakobson;& Cuddy, 2000)**.** Kilgour et al. (2000) investigated in several experiments the influence of musical training and presentation mode (sung condition and spoken condition) on participants' recall ability. The first experiment showed that recall was better for the sung condition, and music training did not have any significant effect. Kilgour et al. (2000) suggested that there might be other characteristics explaining differences in recall between spoken and sung condition, for example, the overall tempo of presentation that was usually slower in the sung condition compared to the spoken one. During the second and third experiments, Kilgour et al. (2000) again looked at the effect of presentation rate on recall, with the duration of both conditions equated. The results were different to the first experiment: participants exposed to the spoken representation showed better recall than the sung condition respondents. These results supported the idea that slower tempo of sung representation could have lead to easier and better recall. In the third experiment the influence of different tempo rates was also investigated. Two different conditions were used: slow tempo samples (30 beats/min.), and a fast tempo samples (70 beats/min.). Once again, results showed that there was some advantage, though not very significant, for the spoken condition.

All this may lead to a conclusion that some benefits of the sung condition comparing to the spoken one revealed in previous studies might actually be caused by the difference in presentation rates (more precisely, tempo rate) for sung and spoken conditions, because usually sung conditions tend to be presented in a slower tempo (Kilgour et al., 2000), giving more time for processing the information, and therefore helping memorization.

In the study provided by Racette and Perets (2007), university students were asked to learn and then perform an unfamiliar song in three conditions (sung-sung, sung–spoken and divided–spoken). An advantage for word recall in the sung–sung condition had been predicted, but results demonstrated the evidence against this hypothesis: fewer words were remembered in the singing condition, both in short and in long-term recall (Racette & Peretz,

2007). Therefore, this data may imply that text and melody are separately represented in memory, which doesn't facilitate the learning process, making singing a kind of a dual task.

Racette and Peretz (2007), analyzing in their article previous studies, say that, though testing the idea of music facilitating text recall requires consideration of both input (perception of the sung text) and output (recall) factors, music's impact for word recall starts at the encoding stage. Thus, the study expresses the idea that words sung are easier to encode than words being spoken. Nonetheless, supporting data for this is mixed. Some of the experiments showed the same or even worse results for sung text comparing to spoken; while many others presented an advantage of sung over spoken presentation. Appealing to Kilgour et al (2000) and Wallace (1994), Racette and Peretz (2007) mentioned that this encoding advantage of sung over spoken text is attributed either to speed or to melody simplicity. Lyrics sung to a complex or changing melody can be even more difficult to remember than their spoken version. The study also remarks that songs possess structural characteristics that may assist text perception and recalling, for example, the metrical structure of music and the number of musical notes in a line can facilitate recalling words. But it is possible, that it is only motivational and emotional aspect of the music that may influence the learning process (Racette & Peretz, 2007).

2.3 Music and Language – neurophysiologic origins

As we can see, data concerning using music in the educational setting is mixed. While most of the studies provide evidence about its positive influence, there is also some contradicting information. As it was mentioned above, the effect of music on word recall begins at the encoding stage. Hence, there may be a reason to overview studies on music and language processing. Investigations in the physiological area can provide some support for the idea of relationship between music and language.

One of possible ways to study neural bases of language and music is through examining songs, which are a kind of exclusive combination of these two cognitive areas, united in one two-dimensional acoustic signal. By studying physiological relationship between music and language and using linguistic or musical components of songs it is possible to gain important data about neural networks underlying language and music cognition. In songs there are prosodic features of speech along with musical melody, which also makes songs an excellent domain for studying the relationship of music and language. For example, both language and music have their own metrical structures. It makes songs an

ideal medium for examining the role of rhythm and meter, while comparing timing and accentuation in speech and music. One of the crucial points of the discussion about interaction between music and language in the form of songs is how melody and text are assumed to operate. If there is an integrated processing, the melody of a song may reinstate the text and vice versa. If melody and text are operated separately, the melody may or may not have connection with words. So, an integrated way of song processing would provide some facilitation effect for remembering and recalling songs vs. spoken text, while separate-memory processing would not.

Investigation of music-language connection and interaction is one of the areas where scientific and humanistic knowledge can closely collaborate, revealing new interactions between traditional boundaries and finding new ideas. This interdisciplinary approach in studying music and language gives a powerful way to investigate mechanisms of sound production and perception, providing both practical and theoretical knowledge about music and language. In the past few years, investigations in the area of different fields of music processing and their neural correlates have significantly progressed. Many experiments have been conducted on syntactic processing in language and music, to determine what is common to syntactic processing in language and music and what kind of effect music has in general learning processes. It is found that a proper combination of music and words facilitates word segmentation, recognition and remembering. There were also studies on specific music parameters that might have influence this area, such as spectral information, temporal structure, melodic and harmonic structure and so on.

For example, Koelsch, & Siebel (2005) have provided an overview of current studies showing that neural networks of language and music perception are partly overlapping. They also pointed out that music perception included complex brain functions underlying “acoustic analysis, auditory memory, auditory scene analysis and processing of musical syntax and semantics”. Furthermore, music perception could influence emotion, autonomic nervous system, hormonal, and immune systems (Koelsch & Siebel, 2005).

There is also other evidence that may explain the positive influence of music on learning and recall of information. For example, studies conducted by Serafine, Crowder & Repp (1984) and Serafine, Davidson, Crowder & Repp (1986) have revealed that music and words are integrated in memory even in the case of senseless syllables**.** Another study, performed by Crowder, Serafine & Repp (1990) showed that music and text served as cues to each other, because of physical interactions or “association by contiguity” (Serafine, Crowder

& Repp, 1984; Serafine, Davidson, Crowder & Repp, 1986; Crowder, Serafine & Repp, 1990).

Samson and Zatorre (1991) fulfilled an experiment investigating dual encoding for songs (in particular, neural mechanisms underlying multiple encoding of songs were examined). In this experiment participants with some lesions in the right or left temporal lobe were involved. Results showed that in the process of words recognition the left temporal lobe was mostly used, but in the case of recognition of melodies both the right and left temporal lobes were involved (Samson & Zatorre, 1991). Such a difference in roles of both lobes provided evidence for existence of dual memory codes. In the case of dual coding melody cues for lyrics could lead to easier text recall.

Schön et al. in their study "Musical and linguistic processing in song perception" presented a short overview of the behavioral, electrophysiological, and neuroimaging research on functional and structural interactions of music and language (Schön, Leigh Gordon & Besson, 2005). Another example in this area is the research, conducted by Brown and colleagues (2006), "Music and language side by side in the brain: a PET study of the generation of melodies and sentences" (Brown, Martinez & Parsons, 2006). Both of these studies showed a clear overlapping activation in brain parts relevant to music and language.

There are a plenty of studies claiming that language and music have a number of similarities on many levels, like sound or structure. Music and language are also connected in terms of general domain properties. Fedorenko et al. (2009), appealing to Patel, 2008 and Bernstein, 1976, explaining this statement, tell that both language and music have a salient rhythmic and melodic structure, both language and music are rule-based systems, that use basic elements like words or tones to create a number of higher order structures (sentences or harmonic sequences). Fedorenko et al (2009), by manipulating sung stimuli, studied the relationship between language and music in a self-paced listening paradigm, trying to reveal, whether language and music have common cognitive resources for structural processing. The study showed an interaction between linguistic and music perception, providing data for existence of an overlap in structural processing of language and music (Fedorenko, Patel, Casasanto, Winawer & Gibson, 2009).

Koelsch et al., investigating simultaneous processing of language and music by means of visually presented sentences and chord sequences, showed how processing of musical syntax interrelated with processing of linguistic syntax. The results spotlighted a clear

overlap of neural resources used in processing of syntax in music and language (Koelsch, Gunter, Wittfoth & Sammler, 2005).

Maess, Koelsch, Gunter and Friederici (2001) studied the neural substrates, which processed music-syntactic incongruities, by means of magnetoencephalography (MEG). The results showed activation of the Broca's area and its right-hemisphere homologue, and that these areas seemed to be involved both into syntactic analysis during auditory language comprehension and into analysis of incoming harmonic sequences. These results allowed to conclude that brain regions processing syntactic information were less language-specific than it had been supposed before (Maess, Koelsch, Gynter & Frederici, 2001).

Besson and Schön (2003) investigating similarities and differences of language and music from the evolutionary and the cognitive points of view, conducted several experiments to evaluate some levels of processing in language and music. This study supported the idea that both in language and music processing general cognitive principles were involved, and analysis of the temporal structure showed same effects in language and music (Besson & Schön 2003).

Two experiments on song perception by Schön (2010) were designed to investigate the domain specificity of linguistic and musical processing. The gained data provided clear evidence about interactions between linguistic and musical elements, giving additional support for the view that shared cerebral network was used for both lexical/phonological and melodic processing (Schön et al., 2010).

Therefore, as we can see, there is a lot of evidence that music processing has shared functions with language processing, providing support to the idea of using music in the learning setting. However, some contradicting data can be found as well. Some studies claim that melodies and words are processed independently: while listening to a song, participant divides attention between text and tune, and these processes do not use the same resources (Bonnel, Faita, Peretz & Besson, 2001). This may lead to a suggestion that song is not a single two-dimensional memory representation, but rather two separate memory representations with one dimension each (Bonnel et al., 2001).

Magne et al. (2004) conducted an interdisciplinary research on rhythm processing in music and language, discussing general aspects of rhythm and the interaction (perception of rhythmic and semantic violations) between language and music. It was shown that rhythm processing might be obligatory in the process of melodic sequences perception, but with the linguistic information processing it seemed to be modulated by attention (Magne et al., 2004).

Ystad (2007) conducted a study similar to those designed by Magne et al. (2004), investigating similarities and differences in meter/rhythm or semantics/harmony perception. It was noticed that data implied different rules for rhythmic modifications processing in music and language (Ystad et al., 2007).

Maidhof and Koelsch (2010), when studying the influence of auditory selective attention on processing the syntactic information in music and language, were not able to find clear evidence concerning interaction of neural resources for syntactic processing (Maidhof & Koelsch, 2010).

Grimshaw & Yelle noted that previous research had found affective prosody to contribute speech perception (usually corresponding with left hemisphere), involving the right hemisphere into the language processing. Grimshaw and Yelle tested the idea, whether melody could demonstrate similar effect. However, results didn't provide evidence for this hypotheses, showing no advantage for the sung text and suggesting that melody didn't facilitate right hemisphere linguistic processing (Grimshaw & Yelle, 2008).

Besson and Schön (2000) reviewed a number of language and music processing studies and found that several important language areas in brain were also involved into music processing, while other features were processed differently (for example, processing of pitch and beat in music and language).

This contradiction between studies may be explained by the Brown's (2001) finding. He introduced a "musilanguage" system, claiming music and language being homologous and having shared and parallel functions. According to him, these functions are of the same origin, and these shared/ parallel features have evolved prior to the distinct, domain-specific features. When speaking about evolutionary connection between this two domains (music and language), it may be helpful to distinct three different types of features. Second step is to discuss models for their respective brain localizations. Thus, music and language have 1) shared features (that are identical for music and language), 2) parallel features (analogous, but not identical), and 3) distinct features (specific to each domain) (Brown, 2001). The author also introduced discrimination between these three types of features and proposed a model for instantiating them into modern brain. These shared features include general processes of vocalization or affective prosody and processes mediated by shared modules. Here also belong expressions of emotional states in music or language. Some parallel features like discreteness, phrase formation and phrasing are mediated by duplicate modules. And finally,

such features as music's isometric rhythms and pitch and language's use of words and propositional syntax are distinct and mediated by diverse neural areas (Brown, 2001).

The idea of these three types of modules itself implies that during the divergence of music and language from the "musilanguage" origin, shared functions came to adopt the same neural areas, parallel functions came to occupy more-or-less corresponding areas in both hemispheres, and localization of the distinct features took place in diversified arrangements.

Within the field of neurophysiological studies, it is possible to find examples of each of these arrangements. When speaking about shared features, the neural substrates for vocalizing and reading both in music and language seem to significantly overlap. The evidence for duplicate modules corresponds to the localizations of musical and language functions in such areas of brain as superior temporal cortex and inferior frontal cortex. And finally, meter, absolute pitch in music and word lexicons, propositional syntax in language are domain-specific and demonstrate diversity of arrangements quite different from those connected to the shared and parallel features (Brown, 2001).

The whole theory about this "trichotomy" of cognitive features in music and language, while highly speculative, can be quite helpful in creating neuroimaging experiments. The idea that music and language are homologues may clarify a lot of similarities and differences of these human-specific forms of auditory communication. Of course, some further explorations of the common and distinct characteristics of language and music is required, as well as exploration of the brain areas used for their processing and production, but it is clear that there is a relationship to be investigated.

2.4 Text and melody – what is in common, what is different?

Concerning the positive influence music may have on the language learning process, another question occurs: how music can contribute word recalling, recognition and segmentation? What features music has to facilitate this process? Do music and language have many characteristics in common? According to Wolfe, both music and speech perceive acoustical features categorically. In both cases many acoustical features are similar and they are perceived and analyzed by the same organs, although used in different ways. However, the process of encoding different elements using these features differs in music and speech. In terms of acoustics, music and speech are basically similar. On the other hand, speech and music are different functionally, because unlike information

is encoded in fundamentally different ways (Wolfe, 2002). So, how can we manipulate with this similarity and dissimilarity to make the process of word perception and remembering easier and, for example, to reduce the number of mistakes made due to inappropriate segmentation? How music can help with it?

2.4.1 Recalling

Serafine, Crowder and Repp conducted a series of experiments to investigate the interaction between melody and text recalling. Results revealed that listeners recognized melody better in presence of text (irrespective of the meaning of the text) and vice versa. This phenomenon was called "integration effect". Serafine and colleagues gave a couple of explanations for this effect: the physical-interaction hypothesis (one element of a song exerting subtle but memorable physical changes on the other element), and association-by-contiguity hypothesis (two components processed in close temporal proximity becoming associated in the memory in a way that each acts as a recall cue for another). Results of the studies provided support for both of these explanations (Crowder, Serafine & Repp, 1990).

The results of the experiments allowed also imply in the interaction between melody and text an asymmetrical integration effect showing that, in the process of recalling, song information melody was more dependent on words than words on melody. This effect was also investigated by Nakada and Abe (2005). They looked at the role two elements of melody - rhythm and pitch patterns – played for the text-melody asymmetrical integration in song perception. Their results suggested that melody processing included fundamentally independent processing of rhythm and pitch (rhythm and pitch may correspond to functionally independent domains) and both rhythm and pitch patterns took part in the text-melody asymmetrical integration effect in the process of song perception and recognition (Nakada & Abe, 2005).

Ginsburg and Sloboda (2007) studied the relationship between words and music. In theirs experiment singers were asked to sing an unaccompanied song by heart. In one case words and melody were learnt separately, in another - together. Results showed that participants with high level of musical expertise demonstrated more clear and smooth performances than participants with lower level of musical expertise. This study also confirmed that music in song provided kind of a framework for text recall, and, on the other side, words could give cues for recalling the melody, while recall of one component — words or melody — affected recall for the other, but both these elements were not integrated to the

extent that failure to remember one correctly always resulted in failure to recall another. Concluding, Ginsburg and Sloboda suggested that learning together words and melody might be a successful strategy (however, maybe, only for people with appropriate level of expertise (Ginsborg & Sloboda, 2007).

2.4.2 Segmentation

Every foreign language, when listened to for the first time, sounds like a continuous flow of meaningless sounds, especially due to the fact that usually word boundaries are not highlighted by consistent acoustical cues, for example, pauses or accents. Of course, even if there existed clear acoustical cues to word boundaries, an obvious lack of lexical knowledge would prevent their efficient use.

At first sight, songs may help in speech segmentation in many ways. Not even mentioning emotional aspects of songs that may significantly raise the level of arousal and attention, pitch contours of songs (from the perceptual point of view) can improve phonological discrimination, because syllable changes are usually accompanied by pitch changes. And finally, constant mapping of musical and linguistic structure can enhance learning mechanisms' operation.

Existing studies in this area are quite limited. Mostly, they investigate connections between music and language, comparing some of the characteristics shared by both these rich and highly-structured instances processed by human brain. In any way, some of these studies are closely connected with the questions we are interested in, and other might contribute to our research because they're investigating some musical features and processes essential for word recognition. Now we will consider them more in details.

First of all, let's look at some important characteristics that music and language both share. When speaking about language and music, it has to be said that their elements are hierarchically organized according to certain principles or combinations. Knowledge (or awareness) of these principles develops expectancies according to the previous context. It also affects processing of the coming linguistic or musical events, and provides cues that may be important for the word recognition and segmentation process. Looking at some cues, we can suppose that some non-linguistic information (particularly, musical) may be used for more efficient cues extraction that leads to easier word segmentation.

2.4.3 Prosody as an essential element in the music and language perception

One of the essential elements is, for example, prosody — an abstract, rule-governed level of structure. We can also say that prosody is a complex system of intonation, rhythm and stress patterns. In the linguistic domain, it structures the language on the word, sentence and discourse levels using variations in different acoustic parameters, some of which are fundamental frequency, timing and intensity. In the music, this term refers to manipulations of such sound properties as frequency, amplitude, time and timbre. Palmer and Hutchins (2006), in their paper "What is prosody. Psychology of Learning and Motivation", consider characteristics of the music prosody. The questions are: whether prosody in music is a complex, rule-governed form of auditory stimulation? Do listeners have a systematical response to it? For this purpose, some functions of musical prosody were reviewed. Some of them were a continuous acoustic stream segmenting into its component units, focus and prominence of items, and also coordination of producers and attributing emotional states to producers. The role that musical prosody may play in the learning process was also discussed. The results showed that prosody aided perceptual learning of primitive units. It also provided low-level cues to help segmentation and learning of hierarchical relationships (Palmer & Hutchins, 2006).

As it was shown earlier, there is some contradiction and inconsistency between studies on comparison of cognitive processes connected to language and music. Some studies claim that language and music have separate processing modules. At the same time, many studies have found evidence that music and language do have resources (cognitive and neural) in common. Comparing musical melody to linguistic prosody, it is also possible to find many similarities. As it has been said, music and language share the same essential acoustic features (pitch, rhythm and accentuation). Many studies investigate the relationship between the language and music prosody. Some of them add facts to the general hypothesis claiming that good "alignment" of prosodic and melodic accents in songs facilitates the process of semantic integration. The results also reveal a neural basis for the song perception, as it was already mentioned in the section on neuroimaginery studies. It has also been studied, whether one central mechanism is responsible for the rhythm processing in both language and music.

In their work "Songs as an aid for language acquisition", Schön and colleagues combined linguistic and musical information. They compared language learning based on speech sequences to language learning based on sung sequences. Hypothesis was that

consistent mapping of linguistic and musical information might enhance learning. This hypothesis was confirmed by results, revealing a strong perceptual facilitation of songs compared to speech. Most importantly, this study showed that the process of segmentation of new words may be significantly enhanced by the structuring and motivating properties of the songs' music (Schön, Boyer, Moreno, Besson, Peretz & Kolinsky, 2008). This study is closely connected with our point of interest, because it investigates the role of musical information as a kind of "prosodic cue". That's why we are going to consider it in more detail.

It has to be said again that both, music and language, are organized hierarchically according to certain principles or combinations. Knowledge (or awareness) of this principles permits to develop expectancies according to the previous context. It also affects processing of the coming linguistic or musical events that might be important for the segmentation process. One of the cues most discussed in terms of speech perception and word segmentation is some sort of distributional statistics for sub-word units, or transitional probabilities.

Saffran, Aslin and Newport (1996), in their work "Statistical learning by 8-month-old infants", showed that infants used statistical properties of syllable sequences for extracting words from the continuous speech. In another work, "Abstract statistical learning of tone sequences by human infants and adults", Saffran, Johnson, Aslin and Newport (1999) demonstrated that a similar learning mechanism applies to musical stimuli.

Schön et al, in their work "Songs as an aid for language acquisition", tested the hypothesis that adding musical information, as a form of a prosodic cue, to speech sequences would enhance segmentation process. Word learning and word segmentation based on speech sequences were compared to segmentation based on sung sequences. As it was pointed before, results confirmed the hypothesis, because an essential learning facilitation of songs compared to speech was revealed (Schön, Boyer, Moreno, Besson, Peretz & Kolinsky, 2008). These results showed that the second language acquisition process, particularly in terms of word segmentation, can largely benefit from motivational and structuring characteristics of songs.

In the experiment, conducted by Schön et al., participants listened to 7 min of speech. The hypothesis was that this time might interfere with learning from speech sequences, but would be adequate for learning from sung sequences. Four consonants and three vowels were used to create eleven syllables, after that syllables were combined to six trisyllabic words. These words were organized into monotone and continuous stream of speech without any acoustic cues at word boundaries. In the learning phase participants were

asked to listen carefully to the stream of syllables (spoken or sung) without trying to analyze them, and after that (in the testing phase), they were told to show, by pressing a button, which of two strings might be a word. Results of the Experiment 1 (spoken condition) revealed that participants were not able to separate words from part-words (48% correct, $p = 0.45$). The next experiment (Experiment 2) was nearly identical to the previous one with one exception: the syllables had been sung and each syllable had been combined with a distinct tone, so that each word always had the same melodic contour. The synthesized syllable stream was identical to the first experiment, except for added precise pitch information for each syllable. Results showed that, in the sung condition, participants were able to learn the words (64% correct, $p < 0.0001$), and the ontroduction of music allowed participants to distinguish words from part-words.

So, the question is why and how can the language acquisition process benefit from additional music information? There can be at least three possible explanations. First, music may generally increase the level of arousal or attention that might enhance also the overall performance. Second, using of tonal and discrete pitch may improve perception of word boundaries and enhance phonological discrimination, as syllables may be discriminated not only because of their phonetic properties, but also because of pitch information and pitch gestalt properties (for example, grouping). And final explanation may be that constant mapping of linguistic and music boundaries may improve global transitional probabilities and increase efficiency of the statistical learning mechanism (Schön, Boyer, Moreno, Besson, Peretz & Kolinsky, 2008).

Another experiment (Experiment 3) was designed to find out, which of possible reasons better explained the effect of facilitation through music. In this experiment, statistical linguistic and musical structures were the same, but not in phase any more (word and pitch boundaries did not take place at the same time). More specifically, the second and third syllables were sung on consistent pitches, but the first one could be sung randomly on six different pitches. Results of this experiment were exactly in between, showing better learning than in the first experiment (56% correct, $p < 0.005$), but worse than in the second. These three experiments make it possible to separate the role of the redundant statistical structure and perceptual saliency in language acquisition.

The finding that results of the experiment with the variable syllable-pitch mapping condition (Exp. 3) were lower than those of experiment with constant syllable-pitch mapping condition (Exp. 2) allows to suggest that superposition of transitional probabilities does play an essential role in the learning process. On the other side, the fact that the performance level

in the syllable-pitch mapping condition (Exp. 3) was higher than in the speech only condition (Exp.1) lets suggest that musical properties also play an important role in learning.

Furthermore, when considering that music provides the same effect as speech prosody, these data would be in line with previous results demonstrating that prosodic information is crucial for words segmentation. Though appropriate prosodic cues, for instance, lengthening, had not been added, it appeared that melodic information might make grouping process easier because of the gestalt properties, and this led to facilitation of speech segmentation process. And another fact is also worth considering: the results may demonstrate that (where multiple statistical cues are used) linguistic cues have priority over musical cues. However, the authors point out that all participants were adult non-musicians, and imply that participants with appropriate level of music expertise or infants might have demonstrated different results.

When speaking about this experiment, it has to be added that in this case it is impossible to say whether learners relied more on music or language, since only language learning was tested. Further studies are required to clarify these results, and to see what kind of music's tonal and/or contour properties can influence segmentation process.

The fact may be taken into account that usually redundant information processing is easier, not only when linguistic and musical information is used, but more generally in other cognitive domains. Some additional explanation is provided by the intersensory redundancy hypothesis by Bahrick & Lickliter (2000). This theory (discussed from the perceptual, cognitive, and social points of view) claims that overlapping information for objects and events presented redundantly arises more attention, attenuates more perceptual differentiation and gives more initial advantage to the perceptual processes than the same information presented unimodally does (Bahrick & Lickliter, 2000).

However, for music and language, it needs to be noticed that songs provide certain overlapping spectral and temporal properties, because in the form of songs music and language share the same modality. This rather unique combination can be more effective than a combination across sensory modalities. More investigations are required to examine, whether the results of this experiment are connected to some specific relations between language and music.

Nevertheless, in general, the results of this experiment let us suggest that sufficient usage of both emotional/arousal and linguistic functions can improve the learning process. Moreover, in the process of foreign language acquisition, especially on the very first stage,

word segmentation can largely benefit from using of music in the form of songs. For example, it is possible to suggest that lullabies and children songs, because of their easy and repetitive composition, might contribute not only to emotional and communicative functions, but can be a tool for the speech processing (Schön, Boyer, Moreno, Besson, Peretz & Kolinsky, 2008).

In this domain, the situation is similar to what already has been discussed. While some studies do argue clear positive influence of music, others don't reveal any significant effect. For instance, study designed by Hom (2009) was conducted to investigate effects of different cues on word segmentation. More specifically, the experiment was intended to find out, whether tonal information could give benefits over what is provided by the regular speech cues.

Participants were asked to listen to a continuous speech stream of four types of pseudo-randomly repeated nonsense words (monotone, prosody-enhanced, tonally-enhanced and tonal-word). On the testing phase, participants had to choose, what of the syllable strings were words from the exposure stream. Results were calculated according to the number of correct responses. The experiment revealed a significant facilitatory effect of the prosodic cues (final vowel lengthening), but no significant effect of the music condition. According to the results, it can be suggested that language-specific cues dominate in the process of word segmentation and musical features do not facilitate perception of word boundaries more than usual speech cues do.

The study failed to replicate previous findings and to find any facilitatory effects of musical elements on word segmentation. According to the author, this might be due to the detrimental effects of two unexpectedly high between-word transitional probabilities. Another explanation might be a possible lack of statistical power. The current results can show only a greater influence of the language-specific prosodic cues in the process of word segmentation. The question of facilitation effect by adding musical information to the speech input still needs to be reinvestigated and accurately measured.

2.4.4 Grouping – current and prospected area of research

Something about grouping must be added to the discussion on the process of word segmentation. People naturally group sounds together into larger rhythmic units. Grouping is an essential feature of the speech and music perception. It affects, for instance, how people break a continuous stream of sounds into words and phrases. The rules of this grouping were

claimed to be universal aspects of auditory perception. Schön et al (2008) suggested that using tonal melody might improve perception of word boundaries and enhance phonological discrimination, because syllables might be discriminated not only because of their phonetic properties, but also by gestalt properties (more precisely, grouping). It was revealed that melodic information might make grouping process easier because of the gestalt properties, and this led to facilitation of the speech segmentation process (Schön, Boyer, Moreno, Besson, Peretz & Kolinsky, 2008).

Patel A.D., in the work "An empirical comparison of rhythm in language and music", observed the overlapping area between musical and speech rhythm in terms of perceptual grouping (mental clustering of events into units) at different hierarchical levels. It seemed that music and speech had shown many similarities (marking group boundaries in similar ways by pitch and duration). According to this, grouping in music may have strong connection with prosodic grouping abilities. These results may be helpful for investigation of the role grouping might play in word segmentation (Patel & Daniele, 2002)

Concerning grouping, also another point of interest occurs. Inversen, Patel and Ohgushi (2006), in their work "How mother tongue influences the musical ear", investigated correlation between language experience and process of grouping in speech and music. This study showed that listeners from the Western and Eastern culture groups perceived simple tone patterns in different ways. For example, they found different rhythmical patterns in identical sound sequences. This difference might be closely related to the rhythms of predominant language. Therefore, it can be suggested that mother tongue affects the way we perceive sounds on a very basic level. Inversen and colleagues found that perception of rhythmic grouping actually varied by culture. An explanation for this difference may originate in speech rhythms. It is suspected that typical rhythmic patterns in the native language might have an influence on rhythm perception in general. Then grouping preferences might be predictable from the structure of small linguistic chunks (Iversen, Patel, & Ohgushi, 2006).

A great deal of the work on rhythmic grouping was done with speakers of Western European languages. These languages have important differences; however, they all put short function words at the onset of small linguistic phrases. It may lead to the some similarity of perceptual grouping in these cultures. Although an analysis of the cultural differences of rhythmic grouping is not in the scope of our paper, it might be an idea for future research.

2.4.5 Pitch contour as a part of melody

Another characteristic of spoken and musical events is contour. Its extraction enables recognition, segmentation and discrimination of short items (both musical and spoken). It has also been noticed that a distinct pitch contour of a melody may help in words perception. Stevens and Keller tested the hypothesis that perceptual biases (for instance, stronger sensitivity to pitch contour in tonal languages) persisted into later auditory processing. These experiments investigated the effect of language background (tonal versus non-tonal) on discrimination of contour using context of speech stimuli, musical intervals and frequency discrimination in Thai and English. Results showed that adult participants with a tonal mother tongue had been more accurate, compared to non-tonal language participants, in the task of discriminating contours in words and speech. The influence of language on discrimination accuracy was revealed for both Thai and English. However, language background showed no influence on discrimination of musical intervals. There wasn't also any evidence of variations in frequency discrimination (Stevens & Keller, 2001).

Another relation between music and speech was investigated by Ross, Choi and Purves in their work "Musical intervals in speech". In this study, database of English vowel phones was analyzed. The aim of the research was to examine the hypothesis of arising musical intervals from the formants in speech. Results showed that the frequency relationships of the first two formants in vowel corresponded to twelve intervals of the chromatic scale. This may explain human preference for the intervals of the chromatic scale and also the relationship between music intervals and speech formants creating phonemes (Ross, Choi & Purves, 2007).

Medeiros (2008) compared sung and spoken sentences to find out how speech intonation influenced song's melody, investigating differences and interactions between spoken and sung intonation components extracted from a Brazilian song. It was revealed that, in terms of pitch contour, musical melody tried to maintain a connection to speech intonation. It was concluded that composers produced (or followed) prosodic rules usual for song competence, not by canceling speech prosodic rules, but rather by transforming them, so as to provide some sort of symbiosis of speech and musical components (Medeiros, 2008).

These studies, though not directly connected with our points of interest, show some similarities and interactions between linguistic and musical elements.

2.4.6 Rhythmical properties and their influence on the song processing

In the discussion on the rhythm as another essential characteristic of music, one of the main questions about musical rhythm is its relation to speech rhythm. Rhythm is widely acknowledged to be an important feature of both speech and music. Nonetheless, there is little empirical data comparing rhythmic organization in these two domains. According to Stevens (2001), one approach to the empirical comparison of rhythm in language and music is to break rhythm down into subcomponents and compare each component across domains. This approach reveals empirical evidence that rhythmic grouping is an area of overlap between language and music, but no empirical support for the long-held notion about a periodic structure in the language comparable to that of music (Stevens & Keller, 2001). Focusing on the statistical patterning of event duration, new evidence suggests that linguistic rhythm of a culture leaves an imprint on its musical rhythm. The latter finding suggests that one effective strategy for comparing rhythm in language and music is to determine, whether differences in linguistic rhythms between cultures are reflected in differences in musical rhythm.

In the psychological research, music and rhythm have been shown to benefit the process of memorization. Memorization seems to be enhanced when various types of verbal information are presented simultaneously with music. Concerning the rhythm, literature also shows that the maximum retentive effect of the rhythm reveals when verbal information makes some sense. It is also remarkable that the biggest impact of the rhythm is when the verbal information is really meaningful. Analyzing previous studies and referring back to Isern (1958) and Bottarri & Evans (1982), Medina (1990) concluded that additional evidence showed that music's benefit was not limited to the rote memorization process (Medina, 1990).

Investigating the relation between music and language, Patel & Daniele, in their work "An empirical comparison of rhythm in language and music", discussed the connection between rhythmical properties of a certain language and its influence on the structure of music produced by the country of this language. While musicologists and linguists have used to suggest that the prosody of a particular language can influence the structure of instrumental music, there was an obvious lack of empirical data supporting this idea. Patel and colleagues investigated speech rhythm and compared rhythmic patterns in English and French language and classical music of these countries. It was revealed that both musical tunes and rhythm of spoken English and French differed significantly in terms of rhythm. This result let the authors say that spoken prosody had an obvious impact on music (Patel & Daniele, 2002).

Similar results was revealed by some other studies as well, for example, by Huron and Ollen (2003), comparing not only excerpts of English and French music, but also music from many other countries like Norway, Poland, Russia and Italy. It was shown that the rhythmic structure of a language demonstrated a strong influence on musical composition, underlying obvious similarities between rhythmic production in language and music.

In the case of meter, the situation is quite different. Music with a regular beat (periodic pulse that affords "temporal coordination between performers and elicits a synchronized motor response from listeners") differs according to cultural traditions (Nettl, 2000). Musical beat is strongly connected with meter (hierarchical organization of beats when some beats are perceived stronger than others). Speech also possesses some kind of metrical hierarchy based on stress (Selkirk, 1984). It may lead to a suggestion that organization of rhythmic sequences according to hierarchical prominence patterns may have its origin in the language. Nevertheless, investigation of stressed syllables of speech has not revealed any regular pulse. This is a very important cognitive difference: the use of a perceptually isochronous pulse in music engages periodic temporal expectancies. These expectancies play an essential (or even basic) role in music cognition (Jones & Boltz, 1989). On the other side, they seem to play little or no role in ordinary speech perception (Pitt & Samuel, 1990). Humans can extract periodicities from complex auditory stimuli. They are also able to focus their expectancies on periodicities at different hierarchical levels in music. So, the question is: regarding songs and their properties, can these expectancies facilitate the process of word recognition?

When considering the role of rhythm for speech perception, another question arises. Regarding the rhythm as a regular change of the strong and weak beats (in music) and syllables (in speech), it is assumed that the number and selection of words compatible with it is usually limited. Therefore, in songs, rhythmic structure (particularly rhymes) with a constrained number of syllables might be used as an appropriate format for setting words to tones. Both in songs and verses, recalling an exact stress pattern activates a metrical grid, providing some cues for more sufficient word recalling. By means of clear metrical structure, words in songs and verses are organized in a common hierarchical structure, thereby helping the memorization process.

Racette and Peretz pointed out that in songs lyrics have some advantage for the process of recalling (Racette & Peretz, 2007). Such a phenomenon of perception of words and melody was described in a number of studies. Firstly, it should be mentioned that lyrics, when organized in a poem, provide some benefit for word recall because of using several linguistic

cues (such as semantics, rhymes and line structure) that facilitate remembering. Rubin (1995), in the book "Memory in oral traditions: The cognitive psychology of counting-out rhymes, ballads, and epics", noted that "repeating patterns of sound in the form of rhyme and alliteration cue memory more broadly and in less time than either imagery or meaning" (Rubin, 1995). So, the question is: when songs do have positive influence on word perception and remembering, what are features that really facilitate these processes. Are these speech cues, like rhyme or line structure? Or musical rhythm, line structure and pitch accents of the melody are also effective in facilitating lyrics recall?

2.5 Music and phoneme perception

The last question that fits into the area of our interest is the influence of musical properties on the phonetic image perception. For our purposes, it would be quite helpful to look at possible interaction between phonemes in speech and musical components in melody.

For example, Kolinsky and colleagues conducted five experiments to find out, whether lyrics and melody, as two dimensions of a song, are processed independently or in an integrated way (Kolinsky, Lidji, Peretz, Besson & Morais, 2009). Having reviewed a number of studies, they made a conclusion that for most consonants rapidly changing acoustic information and acoustic cues of formant transitions were typical. On the other side, for vowels, the relationship between more steady-state frequency and stable spectral information was typical, that made them more appropriate to carry melodic and prosodic cues than consonants. From the physiological point of view, it might be explained by hemispherical differentiation – processing of consonants is more left-lateralized than processing of vowels. Appealing to Bonatti at al, Kolynsky (Kolinsky, Lidji, Peretz, Besson & Morais, 2009) claimed that vowels and consonants might also play distinct roles in speech, and because of it humans better perceived non-adjacent regularities based on consonants than on vowels.

In five experiments performed by Kolinsky at al, musically untrained participants were asked to classify bi-syllabic pseudowords sung on two-tone melodic intervals according to the pitch contour, non-word identity, or on the combination of pitch and pseudoword. Results revealed that consonants were processed more independently from melodic information than vowels, and this difference had no connection neither with sonority of phonemes nor with the acoustical correlates of vowel quality and pitch height (Kolinsky, Lidji, Peretz, Besson & Morais, 2009). These results showed stronger processing connection between vocals and melody than between consonant and melody. On the other side,

connections between segmental (phonemes) and suprasegmental (pitch or pitch contour) areas seemed to be modulated by the nature of phonemes. Kolinsky at al, appealing to Melara and Marks, said vowels and consonants to have different relations with the pitch, so that vowels and the pitch might be processed by the same general auditory mechanisms, while consonants were processed on the phonetic level.

Summarizing, the authors claimed that vowels and consonants had different linguistic functions, with consonants being more connected with word identification, while vowels basically contributing to grammar and prosody, which made them more linked to other non-linguistic auditory dimensions, for example, melody.

Another side of interaction between phoneme perception and music was studied by Gromko (2005) who undertook an experimental research to investigate the influence of music instruction on phonemic awareness or, more specifically, phonemic segmentation ability in kindergarten children. Kindergarten children taken from different schools were divided into the treatment and control groups, and in both groups the same amount of reading instruction was used. The only difference was that the treatment group had every week additional 30 minutes of music lessons. Music instruction included singing songs from different cultures and some advanced music methods, like using percussion and kinaesthetic movement or touching graphic charts. All participants were post-tested after about four months of studying. Results showed a significant gain in letter-naming fluency, phoneme segmentation fluency and nonsense-word phoneme segmentation fluency for group with music instruction. It has also to be taken into account that in this experiment there was difference between the two schools chosen for the study (children in the treatment group were from a lower socioeconomic background and showed lower results during pre-testing). There might as well be some differences between classroom teachers, which could have some influence on the final results. But in any way, this study gives some support for the possibility of using music to help with phoneme segmentation development in young readers (Gromko, 2005).

An experiment conducted by Bigland et al. demonstrated that harmonic structure could influence perception of phonemes despite of listeners' level of musical expertise (Bigland, Tillmann, Poulin, Madurell & D'Adamo, 2001). In this experiment, phoneme monitoring was more fast and precise when phoneme was sung with the tonic than with the less stable subdominant chord. This clear interference between semantics and harmony allowed to suggest that music affected semantic priming in song perception.

The results of this study inspired the scientists to go further to discover, whether harmonic structure interacted with processing of semantic information, revealing the influence of music at a higher level of linguistic processing (Poulin-Charronnat, Bigand, Madurell & Peereman, 2005). Their data showed that music did affect semantic component processing, interacting at some stage with phoneme perception. A possible explanation for this data might be found in the Jones' dynamic attention theory (Jones, 1987), claiming that music provided a number of functions to focus listener's attention. Considering western music, the tonic is more referential than, for example, the less stable subdominant, therefore, tonic provides additional cues to attract new attentional resources. This can explain the fact that linguistic processing was performed better on the tonic than on the subdominant.

Summarizing the above, music may affect semantic priming in vocal music, and some musical components (for instance, harmony) may interact with linguistic processing of lyrics exactly as prosodic cues in speech perception (Poulin-Charronnat, Bigand, Madurell & Peereman, 2005).

2.6 Pitch impact on word recognition

Another question we are also interested in is how pitch affects the process of word segmentation. Pitch is a perceptual sound characteristic that can be defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" (American National Standards Institute: www.ansi.org). According to Schön, music pitch allows to define melodic aspects of a musical sequence. Being one of the basic acoustic parameters of a sound, pitch corresponds both with linguistics and music, which becomes especially evident in tonal languages. Then, combined with other acoustic characteristics like duration, intensity and timbre, pitch may serve for expression of an emotional state and contribute to our perception of joy, sadness, anger – both in speech and in music. It's closely connected with the linguistic functions, like segmentation, modality and focus. Also, combined with such rhythmical components like pauses, intonation, accents, it corresponds to prosody in speech. Thus, we can say that pitch as a musical and linguistic parameter lies directly in the area of our interests, because manipulations with pitch are both musically and linguistically relevant.

There are a number of studies investigating pitch perception both in linguistics and in music. Studies comparing these two huge domains also exist. For example, "The music of speech: Music facilitates pitch processing in language" by Schön, Magne and Besson, where they used manipulations of pitch in unfamiliar language. The result of these manipulations

was presented to adults (both musicians and non-musicians) who did not understand this language at all. Results revealed that musicians were able to notice tiny pitch changes better than non-musicians. Also, Wong showed better similarity of the contour of the brainstem reaction to pitch contour of unfamiliar language tones (Mandarin) in musicians compared to non-musicians. These results show that musical expertise influences pitch perception not only in music but also in language and that this influence occurs very early in the auditory pathway, and also that same processes may be responsible for pitch processing in music and in speech (Schön, Magne & Besson, 2004). Another study, "An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies" by Patel (2006), presented quantitative comparisons of spoken and musical pitch contours using a new model of speech intonation perception. This model (prosogram) transforms speech fundamental frequency contour into a series of separate tones and can be used for comparison of speech and music (Patel, 2006).

However, there is an obvious lack of research studying the influence of "overall pitch" (register) of music or language phrase on word segmentation and recognition. The one that may be closely connected with our aims is "Perceptual confusions of high-pitched sung vowels" by Hollien, Mendes-Schwartz and Nielsen. They studied intelligibility of vowels in singing at very high fundamental frequencies. The case, when F_0 (fundamental frequency) was above the region of normal arising of the F_1 (first vowel formant), was mostly considered. The questions were: could such vowels be correctly identified? Second, if they could, would the context provide the necessary information? Were acoustical features also operative? To check this, eighteen professional singers (5 males and 13 females) were chosen. They sung three isolated vowels at high and low pitches at both loud and soft levels, and their singing was recorded. For perceptual purposes, four different types of auditors (professional musical experts, post-graduate students, under-graduate students and non-musicians) were invited to determine identities of these vowels. The nature of confusions with other vowels was also investigated. It was revealed that changes in fundamental frequency had an obvious influence on vowel perception. Other significant observations were that the target tended to alter toward vowels with a first vowel formant just above the sung frequency (Hollien, Mendes-Schwartz & Nielsen, 1999). These results correspond to our hypothesis that the pitch level may influence the overall perception.

The experiment described by Friedrich and colleagues in their work "Pitch modulates lexical identification in spoken word recognition", examined, whether pitch was used for the process of lexical identification in spoken word recognition in stressed languages. Experiment

revealed that pitch, though not as the most important factor, however was efficiently used for spoken word recognition in stress languages (Friedrich, Kotz, Friederici & Alter, 2004).

Lattner, Meyer and Friederici, in their work «Sex, Pitch, and the Right Hemisphere», considered pitch as one of the essential voice characteristics, serving to decode not only linguistic information, but also other parameters, like age, gender and so on. According to acoustic properties, voice information could be analyzed by several acoustic parameters. One of the most important and perceptually relevant parameters was the fundamental frequency (F0) that determined the perceived pitch of a voice (Lattner, Meyer, & Friederici, 2005). So, we can say that pitch as a musical and linguistic parameter lies directly in the area of our interests, because manipulations with pitch are both musically and linguistically relevant. Namely this perceived pitch of a voice, and (more precisely) its influence on word recognition and segmentation is the next area of our interests.

In another work, Johan Sundberg (1987) investigated isolated vowel intelligibility. Results revealed that in high pitches (soprano singing) sung vowels were extremely difficult to distinguish from one another, due to interaction between the vowels' formant frequencies and the resonant frequencies of the vocal tract. It has to be taken into account, however, that female singers, at high pitches, tend to abandon formant frequencies that are typical for normal speech and are extremely significant for vowel intelligibility. So, the question of vowel intelligibility in high-pitched female singing stays acute (Sundberg, 1987).

One of the first researchers who considered this question was the phonetician Stumpf. In his work, "Die Sprachblaute" (Stumpf, 1926), he compared perceived intelligibility of vowels sung by a professional opera singer and two amateur singers. Different vowels were sung at various pitches, and listeners had to try and identify the vowels. Results showed that identification was better when vowels were performed by a professional singer. It was also found that percentage of right identifications dropped for vowels sung at high pitches (for instance G5). We have to add that identification was much better if vowels were preceded by consonants – therefore, for a successful recognition, listeners should demonstrate some lexical transitions (Stumpf, 1926).

Morozov (1965), in his work "Intelligibility in singing as a function of fundamental voice pitch", investigated intelligibility of syllables performed by professional female and male singers. It was revealed that vowels intelligibility lost about 20% correct identifications around the pitch of E4 for male and B4 for female singers. When singing at C5 (male singer)

and C6 intelligibility dropped to 50% correct answers. In the case of female singing right identification dropped to 10% (Morozov, 1965).

Howie and Delattre (1962), in their work "An experimental study of the effect of pitch on the intelligibility of vowels", found that in high-pitched female singing all vowels tended to be perceived as nearly the same. This corresponded to the results got by Scotto Di Carlo and Germain. They used 15 French vowels performed by a professional singer and found that vowels intelligibility decreased dramatically starting from the middle register. The reason might be in singer's jaw position in high-pitched singing that leaded to changing formant frequencies: all high-pitched vowels had been sung with almost the same jaw position, and, therefore, formant frequencies were virtually similar regardless of vowels intended by the singer (Scotto di Carlo & Germain, 1985).

Smith and Scott (1980) investigated the influence of pitch, larynx height and consonantal environment on vowels. Several vowels were performed in four different ways: opera singing, in consonant-vowel-consonant (CVC) context, with a raised larynx and raised larynx with CVC context. Participants had to identify these vowels in randomized sets of ten tokens of each vowel per condition (method of articulation) at each note. Results showed that in high pitch condition (698 Hz) perceived intelligibility of vowels fell to 16%. The mean intelligibility of vowels at the three highest notes (F5, A5, C sharp 6) was 10% for condition 1, 64% for condition 2, 62% for condition 3, and 83% for condition 4. Results also showed that consonantal transitions made vowel identification sufficiently easier. It also appeared that vowel intelligibility varied for different vowels sung at the same pitch (Smith & Scott, 1980).

Perceived intelligibility of high-pitched vowels and syllables may be influenced with different effects. The first is that singers (especially professional) use to systematically change the formant frequency patterns of usual speech that can lead to intelligibility problems. The second reason is that in high-pitched vowels some partials are dispersed over the frequency band that usually provides information used to recognize specific vowels.

Another experiment by Sundberg (1970) investigated the influence of pitch or (more precisely) female singer's deviations from the formant on the vowel intelligibility. In this experiment, several vowels synthesized with formant frequencies remaining always constant at different fundamental frequencies (from 300 to 1000 Hz) were used. Expert listeners tried to recognize each of the given sounds. Results showed that vowel intelligibility decreased as the pitch went up, and, more important, the overall amount of correct vowel recognitions was much lower than in experiments where non-synthetic vowels were used. The reason might be

that singer's articulation facilitated vowel intelligibility. Other differences between synthetic and real vowels or difference in the fundamental frequency might also have influence on this process (Sundberg, 1970).

Sundberg (1970) also studied vowels intelligibility in male singing. It was found that modifications used in singing articulation (lowering of the larynx, for instance) affected some formant frequencies that were very essential for the vowel quality. For example, after measuring frequencies in sung vowels and comparing these with speech frequencies, considerable differences were found.

Dowd and colleagues (1998) used acoustic impedance spectrometer for vocal tract resonances that provided exact real time information about acoustic response of the vocal tract. These results were used in foreign language studies (particularly for pronunciation classes). This work slightly corresponds to our aims, because it found that, in measuring formant frequencies for a high-pitched voice (F0 = 300 Hz), it was more difficult to determine the formants accurately for the "normal" and low-pitched voice (F0 = 100 Hz) (Dowd, Smith & Wolfe, 1998).

On the other side, differences in quality between spoken and sung vowels are well known. Even singers and singing teachers tend to modify one vowel toward another (not replace, but only slightly modify). In this case sung vowels still retain their identity, but the lowest formant frequencies are clearly different.

Summarizing the above, we can say that changes in the "normal speech" formant frequencies modify vowel quality. And this modification can be sufficiently big and interfere with the vowel identity. In this case, consonant transitions become the most important factor for vowel intelligibility. Our aim is to look at how pitch influences phoneme perception if the consonant transitions are present.

As a conclusion, it needs to be said that, though comparison of language and music is widely investigated, there is a clear contradiction in the data about influence of music on speech perception and the learning process in total. There is also a gap of studies in some aspects of segmentation process. We should also add that some existing studies remain unavailable which produces some challenges in investigating this theme. In any way, we hope that our experiment can make some contribution to understanding the role music plays for speech processing.

3 Research questions and hypotheses

This paper deals with multiple questions, answers to which may help to enlighten some aspects of speech segmentation and recalling. It compares sung versus spoken stimuli, as well as stimuli in different pitches. The aim is to analyze listeners' performance based on stimuli conditions: plain spoken, rhythmically organized spoken or sung, recorded in low, high or middle pitch.

One of the problems we're looking at is the process of speech recall, and the first research question can be formulated as: *Does the sung mode facilitate speech recall comparing to plain spoken condition (mode dependency), and, if so, is it due only to language prosody features or musical features as well?* Based on some previous findings and subjective thinking, our hypothesis is that any positive effect of the sung mode that may be noticed in speech recall results would be caused by rhythmic organization of text, and musical features would not add any significant improvement.

Another area of this research is speech segmentation in different conditions, and the second research question: *Is it harder/easier to make segmentation of a sung stimulus comparing to a spoken stimulus (mode dependency), and, if so, is it due only to language prosody features or musical features as well?* Our hypothesis again is that segmentation of the sung stimulus should not be easier, unless it is specially designed for emphasizing and separating words. Any positive effect of the sung mode that may be noticed in speech segmentation results would be caused by rhythmic organization of text, and musical features would not add any significant improvement

The last dimension of the study concerns the tone of stimuli, and is suggested to help in finding, as to whether shifts in voice height (much higher or much lower) with other characteristics left unchanged (e.g. timbre of voice, rate of speech, mode) may influence perception of stimuli. The nature of this influence is also in focus. The research question is therefore the following: *Is stimulus perceived/segmented better or worse depending on its tone (pitch dependency)?* Our belief is that stimuli spoken/sung in the medium pitch may be segmented better, while a change in the pitch may lead to distorted perception of the phonetic image, namely vowel perception. Moreover, we suggest that worst perceived outputs are going to be of the highest tone.

4 Empirical approach

4.1 Method

To find a solution to our research questions, we had to select an appropriate research method. The method chosen for this study was experimental with controlled conditions and a number of variables to be observed. Definition of these variables requires some further discussion.

The aim was to observe the effect of independent variables, such as mode of stimuli (spoken, sung, etc.) and height of tone upon the variables defining ability to recall speech and recognize words. Null hypotheses, suggesting that no such effects would be present, were tested on quantitative data extracted from the outputs collected during the experiment.

While it is more or less clear about *remembering (recall),* where it's possible to count the recalled amount of the original input in some units, the notion of *word recognition* requires further elaboration. As it has already been discussed, by *word recognition* we understand ability to correctly recognize word's phonetic image and find points where one word ends and another starts – word breaks. Therefore, it seemed reasonable to adopt as units of measure the following: syllables (as a measure of quantity, *remembering*), discovered word breaks and correctly recognized phonemes (as measures of quality, *word recognition*).

4.1.1 Variables

On this basis we have developed several countable variables that are shortly presented in this section, while detailed description is to be found in the following chapters.

Independent variables:

- "stimulus mode" – form of the stimulus: sung text, or rhythmically organized spoken text, or plain spoken text,
- "pitch" – height of voice in which the stimulus was recorded,
- and also 2 auxiliary or "control" variables:
    - "set of sentences" – 3 different sets of sentences were generated (for explanations see 4.1.2.2) with different phonetic contents; this

variable allows to check whether phonetic environment should affect possible results for main variables, and

- "sentence number" – the sequential number of a sentence in the recorded track; this variable is introduced to evaluate the risk of responses quality to be affected by the stimuli order.

Dependent variables:

- number of recalled syllables,
- measure of recall accuracy,
- measure of speech segmentation accuracy,
- share of correctly returned vocals in stressed positions,
- share of correctly returned consonants in stressed positions,
- share of correctly returned vocals in unstressed positions,
- share of correctly returned consonants in unstressed positions.

4.1.2 Stimuli

4.1.2.1 Overview

Types of stimuli

As long as we were going to study the effect music components (in the form of songs) might have for speech remembering and segmentation, the stimuli should include both plain spoken samples pronounced in a usual speaking way and samples that resembled songs. This resemblance should at least include some melodic features and rhythmical organization of the text. To avoid any influence (either positive or negative) from other music elements it was decided not to use any kind of music accompaniment. To obtain a clear picture of properties influencing perception (if such influence is found) – "pure" musical features (such as pitch contour) or just the rhythm organizing the text in a certain way – it seemed reasonable to introduce also the third type of stimuli, rhythmically organized samples (or "verse" as they are referred to in tables due to space lack reasons), with the same rhythm as in the sung sample. So, the final decision included three types of stimuli: plain spoken (non-rhythmic) sentence, rhythmically organized spoken sentence ("verse") and sung (rhythmical) sentence. In terms of this paper we speak about "mode" – sung mode, verse mode and plain text mode.

In order to obtain enough data for analysis we definitely needed more than one sample (and probably, more than two) of each kind: plain text, rhythmically organized and sung. A pilot experiment, described later, showed that a set of 12 sentences was too long and made listeners tired; therefore, it was decided to use 9 sentences with 3 of each kind.

Concerning the shifted pitch of samples, the final stimuli were as follows (pitch modulations are described in detail in the section 4.1.2.4 Changing the pitch):

Mode / Pitch	Sung	Verse	Plain spoken
Middle voice	1.1[1]m, 2.1m, 3.1m Original sung rhythmic	1.2m, 2.2m, 3.2m Original rhythmic spoken	1.3m, 2.3m, 3.3m Original plain spoken
High voice	1.1h, 2.1h, 3.1h Original shifted up in pitch	1.2h, 2.2h, 3.2h Original shifted up in pitch	1.3h, 2.3h, 3.3h Original shifted up in pitch
Low voice	1.1l, 2.1l, 3.1l Original shifted down	1.2l, 2.2l, 3.2l Original shifted down	1.3l, 2.3l, 3.3l Original shifted down

Table 1. Pitch and mode of the stimuli

Language of stimuli

The samples were narrated and sung by a female Russian speaker and the language of the stimuli was Russian. The idea for the language choice was that, on one hand, Russian was mother tongue of the speaker (and researchers) which would make the sentences be pronounced and sound natural, and on the other side, the language was not widely known (unlike, for example, English) and it was not likely that potential participants in Finland had been much subject to it.

Length of the sentences

Length of the sentences was chosen with regard to short-term memory capacity. It is widely admitted that humans can hold in STM about 7±2 items for up to 30 seconds (Sternberg & Mio, 2008). These items may be, for example, words or syllables. In our case participants were not familiar with the source language and therefore they couldn't recognize words compounding the sentences. Thus, we accepted syllables as basic units to evaluate recall. Considering the above, the lower border of possible sentence's length was 7 syllables – the amount an average person can remember in one run. But this would be too few for our purposes. On the other side, we were going to repeat each stimulus two times, and this gave us freedom to make the sentences considerably longer.

[1] For labels' meaning see section Ready sentences and Appendix A. Samples

After a couple of pilot experiments (described below) and thorough consideration we stopped on the length of 11 syllables compounded into four 2-syllable and one 3-syllable words with 4 breaks correspondingly. This was the amount that most of participants would manage to remember fairly well without getting confused. Also, the amount gave us quite a good basis for analysis.

Duration of each stimulus was about 2.5 seconds. A challenge was to record all the samples keeping the minimum difference in duration among them. Considering that singing tended to elongate vowels, it was quite hard to stay within 4% of difference between sung and spoken samples with the average of 2.517 sec for plain spoken stimuli, 2.617 sec (0.1 sec or 4% longer) for sung stimuli and the rhythmically organized spoken stimuli being in the middle with the average of 2.567 sec.

4.1.2.2 Linguistic component

Phonemes

For building sentences, we needed first to choose Russian phonemes that shouldn't make difficulties for foreign speakers. For this, we excluded any phoneme that did not exist in the Finnish language (as we expected most of our respondents to be Finns), were hard to recognize and might be hard to transcribe. This primarily concerned consonants, because Russian vowels ([a], [ɛ] (e), [i], [o], [u][2]), excluding [ɨ] (y), were quite familiar for any European ear. The most difficult consonants were: [ʐ] (zh), [z], [j], [t͡s], [tɕ] (ch), [ʂ] (sh), [ɕɕ] (shch). Also, we were very cautious about using such phonemes as [b], [g], [r], [f].

Thus, consonant sounds to be used for sentences construction were: [d], [k], [ɬ] (l), [m], [n], [p], [s], [t], [v], [x] (h). Also, in some cases (in the position before [i] and [ɛ]), palatalized variants of the same consonants were used: [dʲ], [lʲ], [mʲ], [nʲ], [tʲ], [xʲ].

Syllables

Another important matter was to avoid phoneme combinations that might be strange for foreigners because every language had adopted unique rules of phonemes combinability. Especially this concerned consonants combinations, because diphthongs were not present in Russian language and two vowels in a row did not occur often.

[2] Hereinafter we use for sound notation International Phonetic Alphabet symbols with widely used transliterations in brackets when applicable.

To avoid possible problems with syllables definition and sentences construction and possible difficulties for the respondents, it was decided to use only syllables having the following structure: CV, where C stands for consonant and V – for vowel.

Phonetic environment

It is not unknown that the environment a phoneme is put into affects its recognisability, due to different acoustic features it acquires depending on the context (coarticulation phenomena discussed earlier). Thus, as long as we were going to look at phoneme recognition, it seemed crucial to create "comparable" phonetic environments. Nonetheless, developing 9 similarly sounding but still different sentences was, for one thing, hardly possible, and, what's more important, might have stimulated learning effect or misleading effect – both undesirable.

Therefore, 9 sentences were divided into 3 groups ("sets", in terms of this paper) with different phonetic environments. Within the sets, all stressed syllables were repeated, while unstressed syllables might differ from one sentence to another. Some words might occur in two or even all three sentences in a set.

Ready sentences

Each of the described sets contained two rhythmically organized sentences (one of which to be sung) and one sentence in the form of a usual utterance (plain). All of the sentences were grammatically correct and built of existing Russian nouns, verbs and names. However, meanings of most sentences were nonsense.

The table 2 contains the ready sentences transcribed with Latin letters, accompanied by their labels. Hereinafter the following marking is used for sample labeling:

- the first number (e.g. 2 in 2.3) refers the sentence set,
- while the second number (e.g. 3 in 2.3) denotes the mode of the sample:
    - 1 – for sung samples
    - 2 – for verse samples
    - 3 – for plain text samples

Label	Sentence set	Stressed syllables	Duration of the record
1.1	Dáma tóli súku díka padalá		2.55 sec
1.2	Dáli tóni súpu dímu palilá	da to su di la	2.5 sec
1.3	Dána padalá tómi kasú díka		2.9 sec
2.1	Sála dóma nétu típa malakó		2.6 sec
2.2	Sáni dóli nétu tína dalikó	sa do ne ti ko	2.6 sec
2.3	Sámi dalikó dóma putí nétu		2.35 sec
3.1	Mála tími sóku dúri panisú		2.7 sec
3.2	Máni típa kóni dúhi kalisú	ma ti ko du su	2.6 sec
3.3	Mámi palasú kóli vidú tíha		2.3 sec

Table 2. Created sentences

For the rhythmic sentences, the trochee metrical foot was used with the following rhythmic pattern: cVcv_cVcv_cVcv_cVcv_cvcvcV, where "c" stands for a consonant, "V" – for a vowel in the stressed position, "v" – for a vowel in unstressed position, and "_" – for a word break. In our case the sentence included four 2-syllable words, and due to this rhythm stress always came to the first syllable of the word. To eliminate the undesirable effect of learning and mother tongue's influence (in case of Finns, as the Finnish language tends to always stress the first syllable of a word) the last word had stress on the last (third) syllable.

4.1.2.3 Musical component

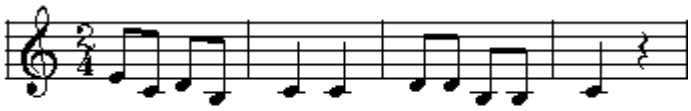
After a couple of trials described below, a compound six-eight meter initially used for recordings appeared to be too complex for our purposes. Three-four compound meter was also put aside, because its application reduced possible number of breaks. So, for our experiment simple two-four meter was found most appropriate.

While creating the melody several conditions were taken into account:

1. Melody should be as simple as possible, to not add any difficulties for processing.
2. Each note should correspond to a particular syllable.
3. Melody was composed in tonal music. Other scales might have been too distracting.
4. Pitch contour should also be simple, without large intervals (larger than fourth) and without skips as well.
5. Preferable intervals were a third and a second.
6. No pauses in sequence should be used, except for the final pause, added to avoid additional vowel lengthening in the end of musical phrase.

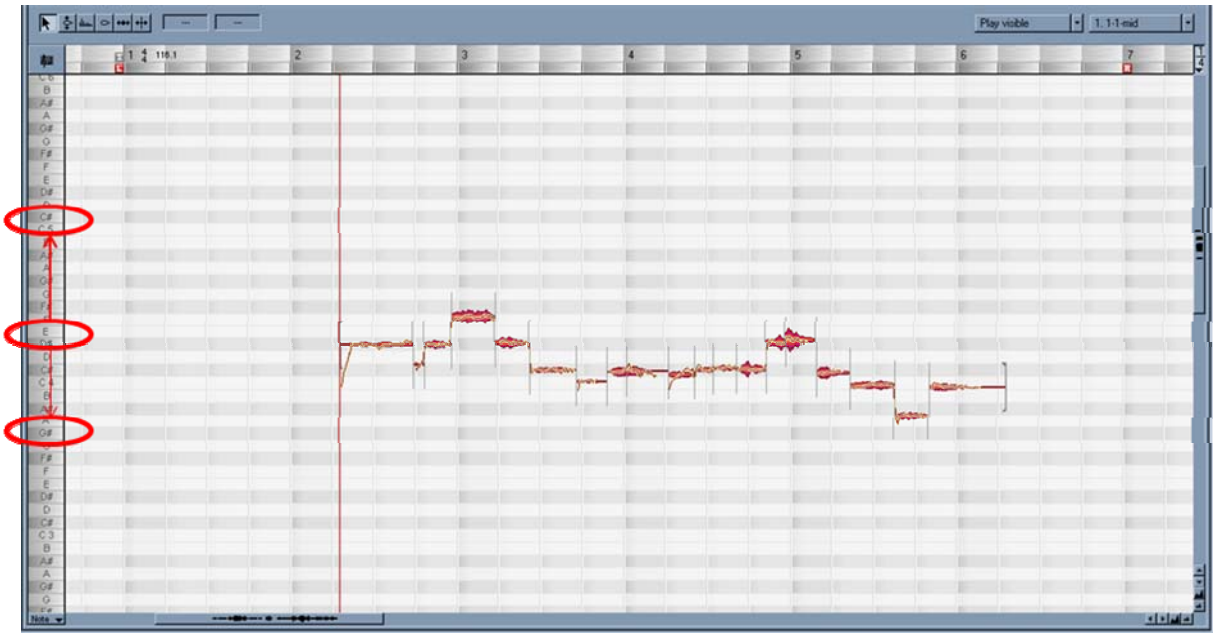
7. Rhythmical structure also should be simple.
8. No dotted notes or off-beats could be used, because they also might have provided additional information to process.
9. The sequence should end with tonic. It was also possible to use referential tonic in the middle of a sequence.
10. The melody should sound naturally, and be easy to perform and to remember.

Notation of the generated melody is presented below:



4.1.2.4 Changing the pitch

To provide the pitch modulation, Melodyne Editor software was used. This application allows manipulating with tonal characteristics by formant shifting. Using the pitch modulation features of the software, the monophonic note sequences were simply moved up (9 semitones) and down (8 semitones). The main criterion for the shift boundary was "naturality" of the sound. This is why the shift up was a bit bigger.



This procedure allowed us to provide the following conditions: 1) all the samples were recorded by the same person, 2) no need to record the same sample several times, 3) the speaker was not required to speak/sing in a higher/lower voice, only in her natural tone, which this let us avoid inevitable articulation variations.

Tonal characteristics of the final samples are presented below:

	Perceptional voice characteristic	Frequency range
Middle voice	Female middle voice	220 – 350 Hz
High voice	Female soprano	370 – 570 Hz
Low voice	Male tenor	140 – 210 Hz

Table 3. Voice tonal characteristics

4.1.3 Pilot experiments

Before the final experiment's design was established, two preliminary experiments had been conducted, to choose proper procedures.

For the first experiment, where 5 respondents participated, 6 16-syllable sentences with 6 word breaks were used. For the sung stimuli an original 6/8-time melody was created. The purpose was to check whether the task of writing down heard stimuli was going to be too complex. The results showed that the procedure was mainly quite comfortable for the respondents. The only problem was an excessive length of the sentences.

After that another trial experiment was conducted with 12 11-syllable sentences. These sentences were presenting two groups: 6 sentences containing 5 words and 4 breaks correspondingly with original 2/4-time melody, and 6 sentences containing 4 words and 3 breaks correspondingly with 3/4-time melody. These groups included samples of different tonalities (higher and lower) and mode (spoken and sung), and different phonetic environments. The purpose was to test two different meters and various phonetic consequences (to see, whether some of them were more difficult for perception). Also, the idea was to test the modulated stimuli to eliminate possible distortions connected with the pitch shift.

Participants (5 students) were asked to write down the samples and answer some questions: "was the task difficult?", "did you find all the speakers similarly easy to perceive?" (by "speakers", different speech tonalities were meant). Nobody from the participants found that voices sounded "unnatural". Another result was choosing the 3/4-meter for the speech and song rhythm. The length of 11 syllables was found suitable for the task.

4.1.4 Design

Nine stimuli described above were recorded using compact custom recorder, processed with PC software and compiled into working tracks. Processing included shifting the samples' pitch up and down using Melodyne software. Denial to use studio equipment for recording was intentional: the record should sounded "natural", without extra clearness, as was usual for most of the records used in language learning classes.

After processing the ready samples were put into 21 quasi-random sequences so that two samples of the same pitch or mode or belonging to the same set would never stand next to each other and the order would always be different. The full list of the compiled tracks is presented on the Appendix A. Samples. This quasi-random order was required to avoid 1) possible "confusion" effect caused by the sentences' order (if participants were, for example, not ready by the start or too tired by the end of listening), and 2) possible learning effect. The general composition of such a track was as follows:

BEEP – **Sent1** – *7s. pause* – **Sent1** – *15s. pause* – BEEP – **Sent2** – *7s. pause* – **Sent2** – etc.

Shorter pauses (7 seconds) were used to distinguish two trials of the same sample, and longer pauses (15 seconds) showed the transition to the next sample. Also, a beep sound came prior to the actual sample, so that the listener could get prepared.

4.1.5 Participants

44 respondents participated in the experiment. The sample was taken from students of the University of Jyväskylä, mostly from the departments of music and philology. It was not representative of the student community, as women number considerably overcame the men number, and the percentage of musicians was bigger than in reality. However, for purposes of our research the group was quite suitable.

Classification of the sample is presented in the tables below. The number of women with musical background was 18 (or 56.3% of the total women number) compared to 8 musically trained men (66.7% of all men), while the amount of women without solid musical education equalled 14 (43.8% of all women) and the corresponding amount of men was 4 (33.3% of all men), resulting in 26 (59.1% in the average) musically trained participants versus 18 (40.9%) of those without musical background.

It was not our target to compare results between genders, that's why the number of females dominated so drastically (by 2.7 times) over the males. The biggest concerns are,

however, connected with the too small amount of musically trained men (only 4, and 2 times less compared to males with musical education).

			Gender		
			female	male	Total
Musical background	solid musical education	Count	18	8	26
		% within Gender	56,3%	66,7%	59,1%
	no musical education	Count	14	4	18
		% within Gender	43,8%	33,3%	40,9%
Total		Count	32	12	44
		% within Gender	100,0%	100,0%	100,0%

Table 4. Musical background / Gender Crosstabulation

In regard to the professional orientation of the participants: for 21 or 47.7% (46.9% of females and 50% of males) of them, the field of specialization was musical, while 17 or 38.6% (50% and 8.3% correspondingly) studied philology (linguistics), and the other 6 participants or 13.6% (with 3.1% of women and 41.7% of men) were into some other fields of studies. This table is presented only to create a clearer picture of the subjects involved, and the data were not actually used to the results' analysis.

			Gender		
			female	male	Total
Speciality	music	Count	15	6	21
		% within Gender	46,9%	50,0%	47,7%
	languages	Count	16	1	17
		% within Gender	50,0%	8,3%	38,6%
	other	Count	1	5	6
		% within Gender	3,1%	41,7%	13,6%
Total		Count	32	12	44
		% within Gender	100,0%	100,0%	100,0%

Table 5. Specialty / Gender Crosstabulation

Other sample representations: the age range in the group was 19 – 30 years with the average of 24 years (however, the age variable was also not considered in the scope of this research); all of the participants spoke at least 3 languages including their mother tongue and none of them had ever studied Russian language.

4.1.6 Procedure and questionnaire

The recorded tracks were copied onto portable players and campus computers. Participants were questioned either individually or in groups of 2 to 9. In either case, each respondent had individual headphones. The intention had been that each one of 21 ready

recordings would be listened to by only one participant. In practice this was generally being observed, some of the tracks, though, were played 2 or 3 times.

The participants were asked to listen to the recording only once and write down, what they heard, using Latin letters. The instructions were included onto the answer form and also given verbally to assure full understanding. After finishing the listening part, the respondents were asked to fill in a questionnaire (on the reverse side of the form). The form with the questionnaire is included into this paper as Appendix B. Questionnaire.

The questionnaire included general data fields, like age, gender, speciality, musical education and spoken languages. The second group of questions concerned the experiment and was intended to find out, what particular aspects of the test had been found the most difficult by the responders. Some of the questions were multiple choices, while also 2 open questions were present.

It took about 10-12 minutes for an average participant to go through the experiment with about 4 minutes of actual listening, about 4 minutes to understand the task and 2-4 minutes for filling in the questionnaire.

Each respondent was assigned a unique number in the format “mus5a” or “non5”, where “mus” and “non” referred to the participant’s background (musical or non-musical), the digit denoted the track number, and the letter “a”-“c” marked cases, where several participants with the same background had been exposed to the same recording.

4.2 Analysis and results

4.2.1 Output data primary processing

The total number of outputs was 393: 44 respondents by 9 samples minus 3 empty entries. Once obtained, the outputs were transferred into Excel spreadsheets like follows:

Track	3.1 high	Mála tími soku dúri panisú		11			4		5	5	5	6	6
Subject ID	**Gender**	**Output**	**Sent #**	**Syll r**	**Syll wr**	**Br orig**	**Brk r**	**Brk wr**	**SSyll r**	**SC r**	**SV r**	**UsC r**	**UsV r**
mus11	f	maile genje soko duri bande so	3	9	2	4	4	1	4	4	3	3	2
mus20	m	mali tuni zu ku tuu pani zu	9	10	0	4	4	2	5	2	3	4	4
mus2	m	male djimi shobju duri bono sju	6	11	0	4	4	1	5	3	5	4	3
mus14a	m	maladimi sookoo duuripani soo	2	11	0	4	2	1	5	4	4	6	5
mus5a	m	maladimi soko duri paniso	5	11	0	4	3	0	5	4	4	6	5
mus5b	f	mala chimi suku … suu	5	7	0	3	3	0	4	3	3	3	3
mus14b	m	doro kuni soopuu	2	5	1	1	1	1	2	2	0	2	1
mus8	f	du ru bandi so	4	5	0	1	1	2	2	2	1	1	3
mus5c	f	mali koni su tu du ri pa ni su	5	9	2	4	4	4	4	4	3	4	4
mus17	f	nalitimasoku duripamiso	1	11	0	4	1	0	5	4	4	5	4
non20	m	male kojo pjietu	9	2	4	1	1	0	1	1	1	1	0
non8	f	malitino soko dulipari su	4	11	0	4	2	1	5	5	5	3	3
non11	f	mala time sooqo duri panyi su	3	11	0	4	4	1	5	5	5	6	5
non17	f	malazimi scniutzuu	1	5	1	2	1	0	3	1	3	2	2
non2	m	mana di la suoku	6	6	0	2	2	1	3	2	3	1	2

Table 6. Output entries presentation

In the table above, the column "Output" presents the actual sentences, as they were written down by participants. The original stimulus is shown in the upper line to make the comparison easier. The column "Sent#" refers to the place the sentence had in the actual recorded sequence. Next columns represent raw data extracted from the outputs (highlighted pink) and include the following:

1. **Syll r** (syllables right) – number of syllables returned "correctly". This point requires further clarification: what we can consider "correct"? Is it an output exactly reproducing the source stimulus? But here we talk about quantity, not quality matters. For example, in the line 4, "maladimi sookoo duuripani soo" it was absolutely obvious that the respondent had written down all the syllables, though he had had perceived some phonemes

distortedly. In another example – line 1, "maile genje soko duri bande so" – it was also quite clear what the participant had meant by "genje", but the phonetic difference was so huge that you'd never have known the word to origin from "timi". Therefore, it didn't seem justified to count these 2 syllables as "correctly" remembered. Not all cases were as evident as these two, but the general rule we adopted for sorting out the outputs was: in terms of recall quantity, a certain item may be regarded as correctly returned, if it is easy (for a native speaker) to make a clear association with the original item.

2. **Syll wr** (syllables wrong) – if an item (syllable) was not clearly associated with any source item, then it was put into this column as wrongly returned syllable.
3. **Br orig** (breaks original) – the number of breaks that corresponded to the returned piece of text. It was introduced to make possible comparing the breaks' number between participants having recalled different amount of syllables.
4. **Br r** (breaks right) – the number of correctly returned word breaks.
5. **Brk wr** (breaks wrong) – the number of unnecessary word breaks returned.
6. **SSyll r** (stressed syllables right) – the number of stressed syllables of the total returned.
7. **SC r** (stressed consonants right) – the number of correctly returned consonant phonemes in stressed position.
8. **SV r** (stressed vowel right) – the number of correctly returned vowel phonemes in stressed position.
9. **UsC r** (unstressed consonants right) – the number of correctly returned consonant phonemes in unstressed position.
10. **UsV r** (unstressed vowel right) – the number of correctly returned vowel phonemes in unstressed position.

4.2.2 Generating countable variables

The raw data presented in the previous section did not allow statistical operations over it. The only exclusion was the quantity data, the amount of syllables. To be able to work with other parameters, we needed to adjust the indicators to a unified basis.

4.2.2.1 F-score

Speech segmentation

To adequately assess the segmentation results, some measure was required that would take into account the amount of breaks returned, the source amount of breaks and the

amount of breaks returned mistakenly. Such a measure might be the F-score ,which is a measure of test's accuracy often used in the field of information retrieval. It considers both the precision and the recall of the test and its general formula is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}},$$

where *precision* refers to accuracy of results and is the number of correct results divided by the number of all results that have been returned; and *recall* represents the share of the source related data that has been returned and equals the number of correct results divided by the number of results that should have been returned.

The results of calculations using this formula lay in the range between 0.00 and 1.00 with 0 being the worst score and 1 – the best value. The traditional F-measure (F_1), when $\beta=1$, is the harmonic mean of precision and recall. But the formula tolerates situations when the number of wrong breaks exceeds the number of right breaks and the value might still equal 1. This was not acceptable for us, because, for our purposes, precision was more important. Therefore, we found it reasonable to use another popular measure, the $F_{0.5}$-score, which weighted precision twice as much as recall.

In our case, the *precision of segmentation* might be represented as the number of correctly returned word breaks divided by the number of all returned breaks, and the *recall of segmentation* – as the number of correctly returned word breaks divided by the number of all breaks existing in the original stimulus. Thus, adjusting the *precision* and *recall* values to the available data, we received the following formula for $F_{0.5}$ (hereinafter referred to as F or F_{br}). Therefore, F_{br} was chosen as the measure for segmentation accuracy.

$$F = (1 + 0.5^2) \cdot \frac{\frac{N_{right}}{N_{returned}} \cdot \frac{N_{right}}{N_{source}}}{(0.5^2 \cdot \frac{N_{right}}{N_{returned}}) + \frac{N_{right}}{N_{source}}}$$

Recall (remembering) accuracy

As long as we wanted to consider not only correctly remembered text, but also the amount of mistakenly returned syllables, we might as well use the F-measure for syllables – F_{syl}. The formula for F_{syl} was the same as shown above, where *precision of remembering (recall)* was the number of correctly returned syllables divided by the number of all returned syllables, and the *recall of remembering* – the number of correctly returned syllables divided by the number of syllables in the original stimulus.

4.2.2.2 Phoneme recognition evaluation

To be able to analyze the correctness of phonetic image perception by participants, we might simply use the percentage measure by dividing the number of correctly returned phonemes by the total amount of syllables returned. Thus, we introduced four variables to assess the perception of consonants and vowels in different positions:

	Consonants	**Vowels**
Stressed position	N_{SC} = SCr / Stressed syll	N_{SV} = SVr / Stressed sylables
Unstressed position	N_{UC} = UsCr / Unstressed syll	N_{UV} = UsVr / Unstressed syllables

4.2.2.3 Variables to be analyzed

Summarizing the said above, the following table presents all the final variables allowing us to analyze the obtained data:

#	Measure	Variable	Formula
	Remembering		
1	Amount of returned syllables	Syll r	= Syll r
2	Accuracy of recall (remembering)	F_{syl}	$=\frac{\left(1+0.5^2\right)\cdot\frac{\text{Syll r}}{\text{Syll r}+\text{Syll wr}}\cdot\frac{\text{Syll r}}{11}}{\left(0.5^2\cdot\frac{\text{Syll r}}{\text{Syll r}+\text{Syll wr}}\right)+\frac{\text{Syll r}}{11}}$
	Word recognition		
3	Accuracy of speech segmentation	F_{br}	$=\frac{\left(1+0.5^2\right)\cdot\frac{\text{Brk r}}{\text{Brk r}+\text{Brk wr}}\cdot\frac{\text{Brk r}}{\text{Br orig}}}{\left(0.5^2\cdot\frac{\text{Brk r}}{\text{Brk r}+\text{Brk wr}}\right)+\frac{\text{Brk r}}{\text{Br orig}}}$
	Correctness of phonetic image		
4	Stressed consonant correctness	N_{SC}	= SCr / SSyllr
5	Stressed vowel correctness	N_{SV}	= SVr / SSyllr
6	Unstressed consonant correctness	N_{UC}	= UsCr / (Syllr – SSyllr)
7	Unstressed vowel correctness	N_{UV}	= UsVr / (Syllr – SSyllr)

Table 7. Countable variables

Note: For the purposes of convenience of calculations we used percentage presentation of the variables 2-7, i.e. multiplying the resulting value by 100.

4.2.3 Data analysis

Statistical analysis of the data was performed using the SPSS v.19 software. The data was checked for outliers and missing values. The following table represents the general descriptive statistics for all 7 variables, where "number of syllables" corresponds the variable **1 "Amount of returned syllables"**, "F-score syllables" is the measure **2 "Accuracy of recall"**, "F-score breaks" – measure **3 "Accuracy of speech segmentation"**, "consonants stressed" refers to the variable **4 "Stressed consonant correctness"**, "vowels stressed" – **5 "Stressed vowel correctness"**, while "consonants unstressed" and "vowels unstressed" represent correspondingly variables **6** and **7**.

The statistical data includes, in particular, the means and standard deviation values for the variables, as well as minimum and maximum values and three percentiles. It is clear from the table, that medians for all the variables are well above the corresponding means, which is usual for non-normal negatively skewed distributions. So, the difference between mean and median for accuracy of recall is 8.2%, for accuracy of speech segmentation – 11.2%, for correctness of stressed consonants recall – 8.1%, for correctness of stressed vowels recall – 15.0%, while for correctness of unstressed consonants recall – 6.3%. And only for the amount of returned syllables and correctness of unstressed vowels recall the difference is 3.4% and 3.0% correspondingly, not resulting in "more normal" distribution, however.

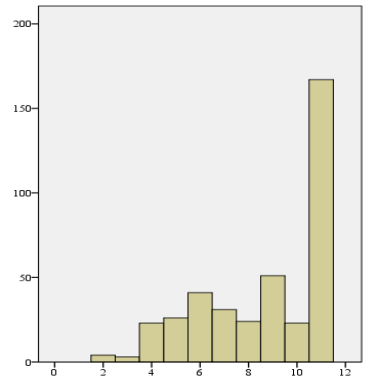
	N	Mean	Std. Deviation	Min	Max	Percentiles 25th	50th (Median)	75th
Number of syllables	393	8,70	2,509	2	11	7,00	9,00	11,00
F-score syllables	393	88,45	14,838	29	100	81,82	95,74	100,00
F-score breaks	393	74,96	25,770	0	100	62,50	83,33	100,00
Consonants stressed	393	73,99	23,771	0	100	60,00	80,00	100,00
Vowels stressed	393	86,93	17,426	0	100	80,00	100,00	100,00
Consonants unstressed	393	78,38	22,262	0	100	66,67	83,33	100,00
Vowels unstressed	393	80,89	21,371	0	100	66,67	83,33	100,00

Table 8. Descriptive Statistics

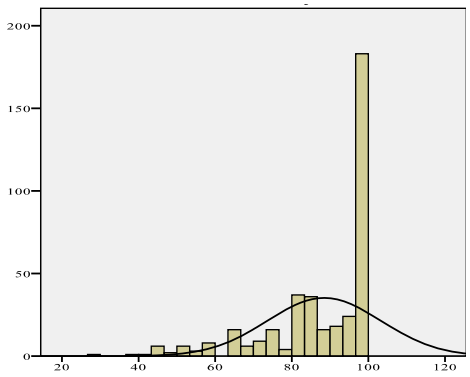
Frequency distribution

The values were screened for normal distribution. The histograms on the Graph 1 show that all the variables were far from normal distribution with frequency curves being J-shaped: The One-Sample Kolmogorov-Smirnov Z test also confirmed that the data for all the observed variables were non-normally distributed: the table below shows that $p < 0.001$ (see below Table 9).

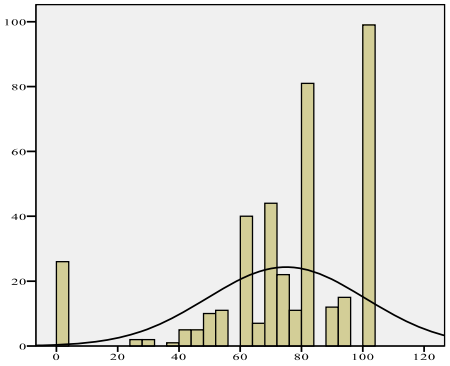
Graph 1. Frequency distributions



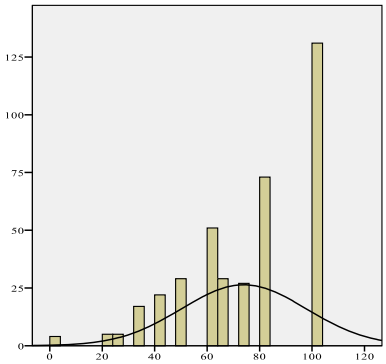
Number of syllables distribution



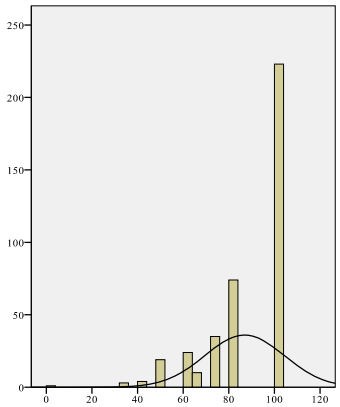
F-score syllables distribution



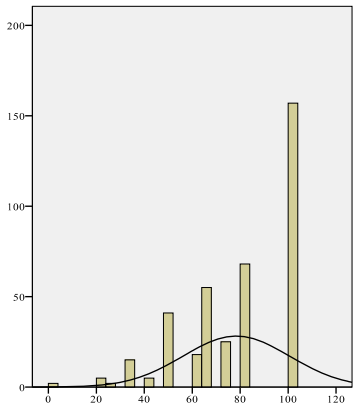
F-score breaks distribution



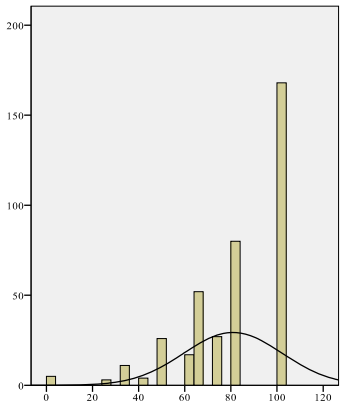
Consonants stressed distribution



Vowels stressed distribution



Consonants unstressed distribution



Vowels unstressed distribution

		Number of syllables	F-score syllables	F-score breaks	Consonants stressed	Vowels stressed	Consonants unstressed	Vowels unstressed
N		393	393	393	393	393	393	393
Normal Parameters[a,b]	Mean	8,70	88,45	74,96	73,99	86,93	78,38	80,89
	Std. Deviation	2,509	14,838	25,770	23,771	17,426	22,262	21,371
Most Extreme Differences	Absolute	,245	,218	,168	,196	,341	,234	,242
	Positive	,180	,218	,166	,137	,227	,166	,186
	Negative	-,245	-,215	-,168	-,196	-,341	-,234	-,242
Kolmogorov-Smirnov Z		4,857	4,325	3,332	3,893	6,758	4,633	4,795
Asymp. Sig. (2-tailed)		,000	,000	,000	,000	,000	,000	,000

a. Test distribution is Normal. b. Calculated from data.

Table 9. One-Sample Kolmogorov-Smirnov Test for all variables

Selecting statistical methods

This kind of distribution meant that mostly participants had been successful and quite efficient in the challenging task of writing foreign sentences down. However, such a distribution made it impossible to use many statistical tools comparing means, like ANOVA test. For our analysis, the mean was not a relevant measure, while the median appeared to be more representative; therefore, nonparametric tests were used for the analysis.

The design of the experiment only allowed a very limited use of related samples tests. For example, it was possible to make comparison between sung-rhythmic-plain stimuli recalled by the same person, because each participant had been exposed to three sets of sentences, so that each sentence of a set had had been recorded with the same pitch, while the difference between sets had been not only in their phonetic structure, but also in the pitch.

Consequently, it was not always feasible to use tests for related samples. In general, the data was treated in such a way, as if each entry had been returned by a different subject, which gave us freedom to apply tests for 2 or more independent samples, like Mann-Whitney U-test and Kruskal-Wallis H-test, which did not concern the distribution pattern and were not sensitive to outliers. However, for the above mentioned mode comparison (sung-rhythmic-plain spoken) the analysis included the Friedman's nonparametric test for related samples.

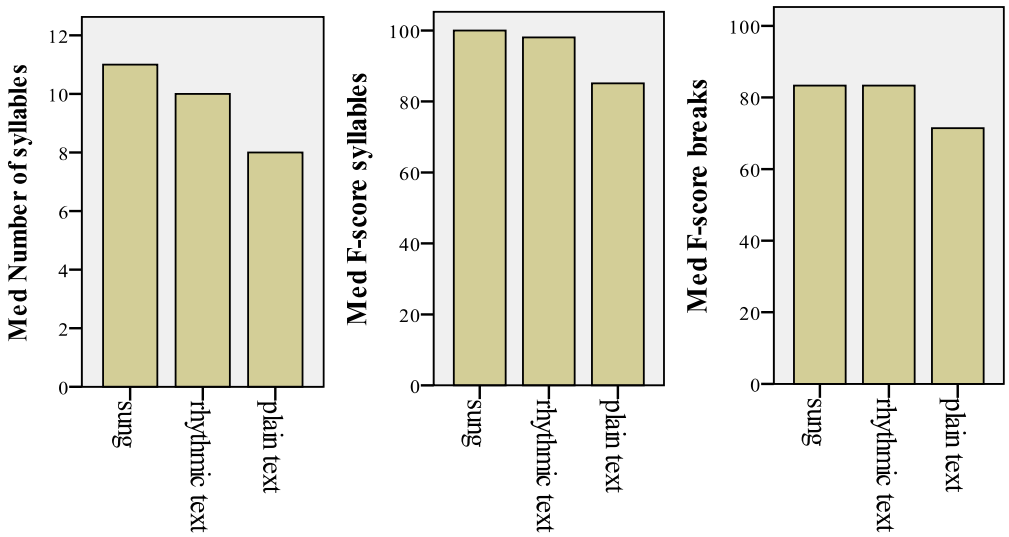
The following chapters describe in detail the analysis process, as well as findings. First, the main issues – mode and voice height – are thoroughly considered, which is followed with "secondary" variables. After that, a quick look is given to intercorrelations between dependent variables.

4.2.4 Results from statistical analyses

4.2.4.1 Influence of mode and voice height

Stimulus mode

Stimulus mode was one of the main conditions to look at within our study. Quite according to our expectations, the analysis showed strong correlations between stimulus mode and variables defining recall and speech segmentation abilities, to be discussed here. The Graph 2 below demonstrates median values relations between different stimulus modes for such variables as amount and accuracy of recall and accuracy of segmentation.



Graph 2. Correlation between mode and recall/segmentation abilities

The first two graphs represent relationships between mode of stimuli and recall ability: amount and accuracy of recall, correspondingly. It is clear from the graph that there was a trend for better recall of the sung stimuli and a definite regress in recall and segmentation of the plain spoken stimuli. Compared medians demonstrate clear distinction between sung/rhythmic spoken stimuli on one side and plain spoken stimuli on the other side. Nonetheless, the difference between the first two modes (sung and rhythmic spoken) was not very obvious. The third graph discovers similar results for the segmentation accuracy variable.

To check the trend, the Friedman's rank test for related samples was applied. For this purpose, the outputs by the same participants were compared within phonetic sets. The test confirmed the tendency, which is clear from the tables below (Table 10): with a very high significance ($p \leq 0.001$), speech mode appeared to affect comprehension and remembering in such aspects as amount of information recalled, accuracy of recall, accuracy of speech segmentation and recognition of vowels in unstressed positions.

Table 10. Friedman's ranking for stimuli modes

Amount of recall, stimuli modes ranks

	Mean Rank
Number of syllables, sung mode	2,30
Number of syllables, rhythmic mode	2,09
Number of syllables, plain mode	1,61
Test Statistics	
N	129
Chi-Square	43,197
df	2
Asymp. Sig.	,000

Accuracy of recall, stimuli modes ranks

	Mean Rank
F-score for syllables, sung mode	2,24
F-score for syllables, rhythmic mode	2,12
F-score for syllables, plain mode	1,63
Test Statistics	
N	129
Chi-Square	34,784
df	2
Asymp. Sig.	,000

Accuracy of segmentation, stimuli modes ranks

	Mean Rank
F-score for breaks, sung mode	2,09
F-score for breaks, rhythmic mode	2,24
F-score for breaks, plain mode	1,67
Test Statistics[a]	
N	129
Chi-Square	28,000
df	2
Asymp. Sig.	,000

Consonants stressed, stimuli modes ranks

	Mean Rank
Stressed consonants, sung mode	2,03
Stressed consonants, rhythmic mode	1,91
Stressed consonants, plain mode	2,06
Test Statistics[a]	
N	129
Chi-Square	2,049
df	2
Asymp. Sig.	,359

Vowels stressed, stimuli modes ranks

	Mean Rank
Stressed vowels, sung mode	1,98
Stressed vowels, rhythmic mode	1,92
Stressed vowels, plain mode	2,10
Test Statistics[a]	
N	129
Chi-Square	3,884
df	2
Asymp. Sig.	,143

Consonants unstressed, stimuli modes ranks

	Mean Rank
Unstressed consonants, sung mode	2,04
Unstressed consonants, rhythm. mode	2,07
Unstressed consonants, plain mode	1,89
Test Statistics	
N	129
Chi-Square	3,176
df	2
Asymp. Sig.	,204

Vowels unstressed, stimuli modes ranks

	Mean Rank
Unstressed vowels, sung mode	2,20
Unstressed vowels, rhythmic mode	2,03
Unstressed vowels, plain mode	1,77
Test Statistics	
N	129
Chi-Square	15,064
df	2
Asymp. Sig.	,001

However, the relations had to be observed more precisely, to be able to make conclusions concerning the nature of the trend: whether it was related to prosody or music features of the stimuli. For this purpose, we applied the Wilcoxon's signed ranks test to compare the modes pairwise. The following table (Table 11) represents the results of this comparison (only for variables that proved to be affected).

For the dimensions of speech recall and segmentation, the results clearly show that plain spoken condition differs from sung / rhythmic spoken conditions with a very high level of significance ($p < 0.001$). Nevertheless, between each other, these 2 latter conditions distinguish significantly only in the respect of remembering ($p < 0.05$), with the median values of 11 syllables (for the sung stimuli) and 10 syllables recalled (for the rhythmic stimuli). This difference in observed values between sung and rhythmic spoken stimuli may partly be explained by variance in duration of the records (about 2%). Therefore, it's hard to

Table 11. Wilcoxon's signed rank test for mode effect

		N	Mean Rank	Sum of Ranks	Z	Asymp. Sig. (2-tailed)
Number of syllables, rhythmic mode - Number of syllables, sung mode	Negative Ranks	45	35,44	1595,00	-2,331[a]	,020
	Positive Ranks	24	34,17	820,00		
	Ties	62				
	Total	131				
Number of syllables, plain spoken mode - Number of syllables, rhythmic mode	Negative Ranks	70	51,06	3574,50	-4,331[a]	,000
	Positive Ranks	27	43,65	1178,50		
	Ties	32				
	Total	129				
Number of syllables, plain spoken mode - Number of syllables, sung mode	Negative Ranks	82	56,63	4644,00	-5,721[a]	,000
	Positive Ranks	24	42,79	1027,00		
	Ties	24				
	Total	130				
F-score for syllables, rhythmic mode - F-score for syllables, sung mode	Negative Ranks	45	38,24	1721,00	-1,563[a]	,118
	Positive Ranks	30	37,63	1129,00		
	Ties	56				
	Total	131				
F-score for syllables, plain spoken mode - F-score for syllables, rhythmic mode	Negative Ranks	77	56,82	4375,50	-4,391[a]	,000
	Positive Ranks	31	48,73	1510,50		
	Ties	21				
	Total	129				
F-score for syllables, plain spoken mode - F-score for syllables, sung mode	Negative Ranks	78	59,46	4637,50	-5,194[a]	,000
	Positive Ranks	30	41,62	1248,50		
	Ties	22				
	Total	130				
F-score for breaks, rhythmic mode - F-score for breaks, sung mode	Negative Ranks	40	52,61	2104,50	-,652[b]	,514
	Positive Ranks	55	44,65	2455,50		
	Ties	36				
	Total	131				
F-score for breaks, plain spoken mode - F-score for breaks, rhythmic mode	Negative Ranks	77	58,31	4490,00	-4,978[a]	,000
	Positive Ranks	30	42,93	1288,00		
	Ties	22				
	Total	129				
F-score for breaks, plain spoken mode - F-score for breaks, sung mode	Negative Ranks	71	58,97	4187,00	-4,726[a]	,000
	Positive Ranks	33	38,58	1273,00		
	Ties	26				
	Total	130				
Unstressed vowels, rhythmic mode - Unstressed vowels, sung mode	Negative Ranks	57	52,54	2994,50	-2,239[a]	,025
	Positive Ranks	40	43,96	1758,50		
	Ties	34				
	Total	131				
Unstressed vowels, plain spoken mode - Unstressed vowels, rhythmic mode	Negative Ranks	61	44,64	2723,00	-1,447[a]	,148
	Positive Ranks	35	55,23	1933,00		
	Ties	33				
	Total	129				
Unstressed vowels, plain spoken mode - Unstressed vowels, sung mode	Negative Ranks	67	56,22	3767,00	-4,047[a]	,000
	Positive Ranks	34	40,71	1384,00		
	Ties	29				
	Total	130				

a. Based on positive ranks. b. Based on negative ranks. c. Wilcoxon Signed Ranks Test

make a definite conclusion about positive influence of the sung mode on speech recall and segmentation. On the other hand, the results absolutely argue importance of prosodic (rhythmic) cues for improving abilities to recall and segment speech.

Another revealed dependency, between the stimuli mode and accuracy of perception of vowels in unstressed positions, showed a quite interesting effect, when unstressed vowels had been much better recognized in the sung mode comparing to the plain spoken mode ($p < 0.001$) and, with a weaker significance, to the verse mode ($p < 0.05$). The difference between two spoken modes (plain spoken and rhythmically organized spoken) was not found to be significant. It is hard to say, what may be the reason for such an effect. One of the suggestions may be that the explanation is not about perception but pronunciation: in singing, unstressed vowels may be involuntarily articulated more precisely than in speaking. However, this consideration requires further studying.

Voice height

This variable didn't allow to apply related sample tests, because stimuli of different pitch belonging to the same phonetic sets had been included into different string sentences (offered to different participants). At the same time, we couldn't compare results between different phonetic sets, as phonetic environment had proved to affect perception. To find out possible dependencies, Kruskal-Wallis test for several unrelated samples was applied to the data. Its results showed that there was no dependency on this condition for any variable, except for the recognition of vowels in stressed positions ($p < 0.001$). Some non-significant trend was also found for perception of consonants in stressed positions ($p = 0.056$) and consonants in unstressed positions ($p = 0.088$).

Pairwise comparisons using the Mann-Whitney rank test revealed that lower pitch (compared to normal) had affected perception of consonants in stressed positions ($p < 0.05$), so that the median value for the lower pitch was 0.75, and for normal pitch – 0.80. While higher pitch had affected (compared to normal pitch) perception of consonants in unstressed position ($p < 0.05$) and (compared to lower and normal) – perception of vowels in stressed position ($p < 0.001$). For this latter aspect, the difference was 0.80 (median value for higher pitch) compared to 1.00 and 1.00 (median values for lower and normal pitch).

Non-significance for other phonemes recognition demonstrates validity of the prepared samples (absence of distortion that could have affected perception). This makes significant relations still more important.

Generally, the revealed dependency supports earlier research, showing that higher pitch affects perception of vowels, especially in singing. To make a closer look at the dependency, we applied the same tests to the data segregated by mode. And these tests showed discrepancy with earlier suggestion that pitch affected only perception of the sung stimuli. The tables below (Table 12 - Table 14) represent the ranking results of the Mann-Whitney test.

Table 12. Mann-Whitney ranks by mode, low voice vs. middle voice

	Voice height	N	Mean Rank	Sum of Ranks	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
Stressed consonants, sung mode	low voice	44	42,43	1867,00	877,000	1867,000	-,780	,435
	middle voice	44	46,57	2049,00				
	Total	88						
Stressed consonants, rhythmic mode	low voice	44	42,00	1848,00	858,000	1848,000	-,760	,447
	middle voice	43	46,05	1980,00				
	Total	87						
Stressed consonants, plain mode	low voice	43	38,07	1637,00	691,000	1637,000	-2,266	,023
	middle voice	44	49,80	2191,00				
	Total	87						
Stressed vowels, sung mode	low voice	44	44,14	1942,00	952,000	1942,000	-,154	,878
	middle voice	44	44,86	1974,00				
	Total	88						
Stressed vowels, rhythmic mode	low voice	44	41,84	1841,00	851,000	1841,000	-,902	,367
	middle voice	43	46,21	1987,00				
	Total	87						
Stressed vowels, plain mode	low voice	43	40,36	1735,50	789,500	1735,500	-1,772	,076
	middle voice	44	47,56	2092,50				
	Total	87						
Unstressed consonants, sung mode	low voice	44	41,47	1824,50	834,500	1824,500	-1,158	,247
	middle voice	44	47,53	2091,50				
	Total	88						
Unstressed consonants, rhythmic mode	low voice	44	46,17	2031,50	850,500	1796,500	-,858	,391
	middle voice	43	41,78	1796,50				
	Total	87						
Unstressed consonants, plain mode	low voice	43	38,08	1637,50	691,500	1637,500	-2,241	,025
	middle voice	44	49,78	2190,50				
	Total	87						
Unstressed vowels, sung mode	low voice	44	39,95	1758,00	768,000	1758,000	-1,801	,072
	middle voice	44	49,05	2158,00				
	Total	88						
Unstressed vowels, rhythmic mode	low voice	44	44,60	1962,50	919,500	1865,500	-,236	,814
	middle voice	43	43,38	1865,50				
	Total	87						
Unstressed vowels, plain mode	low voice	43	39,01	1677,50	731,500	1677,500	-1,879	,060
	middle voice	44	48,88	2150,50				
	Total	87						

Table 13. Mann-Whitney ranks by mode, low voice vs. high voice

	Voice height	N	Mean Rank	Sum of Ranks	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
Stressed consonants, sung mode	low voice	44	44,35	1951,50	961,500	1951,500	-,056	,956
	high voice	44	44,65	1964,50				
	Total	88						
Stressed consonants, rhythmic mode	low voice	44	43,60	1918,50	928,500	1918,500	-,335	,738
	high voice	44	45,40	1997,50				
	Total	88						
Stressed consonants, plain mode	low voice	43	42,40	1823,00	877,000	1823,000	-,421	,674
	high voice	43	44,60	1918,00				
	Total	86						
Stressed vowels, sung mode	low voice	44	50,63	2227,50	698,500	1688,500	-2,425	,015
	high voice	44	38,38	1688,50				
	Total	88						
Stressed vowels, rhythmic mode	low voice	44	48,93	2153,00	773,000	1763,000	-1,721	,085
	high voice	44	40,07	1763,00				
	Total	88						
Stressed vowels, plain mode	low voice	43	49,21	2116,00	679,000	1625,000	-2,341	,019
	high voice	43	37,79	1625,00				
	Total	86						
Unstressed consonants, sung mode	low voice	44	43,15	1898,50	908,500	1898,500	-,512	,609
	high voice	44	45,85	2017,50				
	Total	88						
Unstressed consonants, rhythmic mode	low voice	44	46,88	2062,50	863,500	1853,500	-,920	,357
	high voice	44	42,13	1853,50				
	Total	88						
Unstressed consonants, plain mode	low voice	43	46,13	1983,50	811,500	1757,500	-,992	,321
	high voice	43	40,87	1757,50				
	Total	86						
Unstressed vowels, sung mode	low voice	44	44,72	1967,50	958,500	1948,500	-,083	,934
	high voice	44	44,28	1948,50				
	Total	88						
Unstressed vowels, rhythmic mode	low voice	44	45,66	2009,00	917,000	1907,000	-,450	,653
	high voice	44	43,34	1907,00				
	Total	88						
Unstressed vowels, plain mode	low voice	43	41,70	1793,00	847,000	1793,000	-,681	,496
	high voice	43	45,30	1948,00				
	Total	86						

Table 14. Mann-Whitney ranks by mode, middle voice vs. high voice

	Voice height	N	Mean Rank	Sum of Ranks	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
Stressed consonants, sung mode	middle voice	44	46,22	2033,50	892,500	1882,500	-,649	,516
	high voice	44	42,78	1882,50				
	Total	88						
Stressed consonants, rhythmic mode	middle voice	43	45,13	1940,50	897,500	1887,500	-,419	,675
	high voice	44	42,90	1887,50				
	Total	87						
Stressed consonants, plain mode	middle voice	44	48,74	2144,50	737,500	1683,500	-1,868	,062
	high voice	43	39,15	1683,50				
	Total	87						
Stressed vowels, sung mode	middle voice	44	51,25	2255,00	671,000	1661,000	-2,674	,008
	high voice	44	37,75	1661,00				
	Total	88						
Stressed vowels, rhythmic mode	middle voice	43	50,34	2164,50	673,500	1663,500	-2,491	,013
	high voice	44	37,81	1663,50				
	Total	87						
Stressed vowels, plain mode	middle voice	44	52,91	2328,00	554,000	1500,000	-3,894	,000
	high voice	43	34,88	1500,00				
	Total	87						
Unstressed consonants, sung mode	middle voice	44	46,23	2034,00	892,000	1882,000	-,666	,506
	high voice	44	42,77	1882,00				
	Total	88						
Unstressed consonants, rhythmic mode	middle voice	43	44,02	1893,00	945,000	1935,000	-,009	,993
	high voice	44	43,98	1935,00				
	Total	87						
Unstressed consonants, plain mode	middle voice	44	52,83	2324,50	557,500	1503,500	-3,381	,001
	high voice	43	34,97	1503,50				
	Total	87						
Unstressed vowels, sung mode	middle voice	44	48,73	2144,00	782,000	1772,000	-1,687	,092
	high voice	44	40,27	1772,00				
	Total	88						
Unstressed vowels, rhythmic mode	middle voice	43	44,62	1918,50	919,500	1909,500	-,237	,813
	high voice	44	43,40	1909,50				
	Total	87						
Unstressed vowels, plain mode	middle voice	44	47,16	2075,00	807,000	1753,000	-1,217	,224
	high voice	43	40,77	1753,00				
	Total	87						

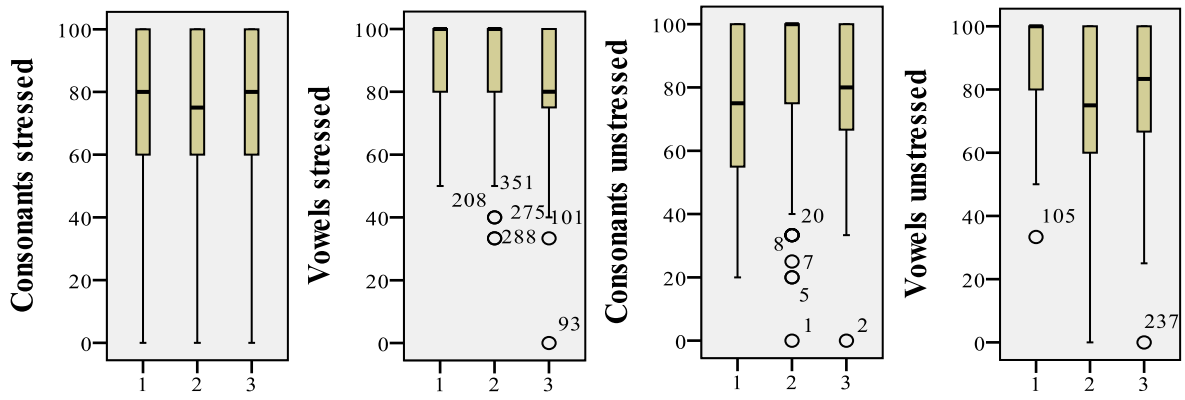
These results indicate that the biggest influence of the voice pitch was on perception of the plain spoken samples. So, the lower pitch (compared to normal) appeared to significantly affect perception of consonants both in stressed and unstressed positions in the plain spoken samples ($p < 0.05$). The same trend was found (slightly above the low significance threshold) for perception of vowels both in stressed and unstressed positions in the plain spoken samples, as well as vowels in unstressed positions in the sung samples. Then, the higher pitch proved to significantly affect perception of vowels in stressed positions in the

sung ($p < 0.01$), rhythmically organized ($p < 0.05$) and plain spoken ($p < 0.001$) modes, as well as perception of consonants in unstressed positions in the plain spoken mode ($p = 0.001$).

From the findings, it is clear that the tendency (high pitch affecting vowel perception) has been much stronger for the plain spoken samples, which contradicts with earlier suggestions this effect to be true only for singing. In regard to the lower pitch, it's hard to make any conclusions, because the recorded lower pitch was within the normal human voice height range. However, the said effect of the higher voice could be taken into account when, for example, preparing language learning materials, because ability to correctly recognize phonemes in a word, or its phonetic image, is essential for speech recognition.

4.2.4.2 Other dependencies

Sentence set



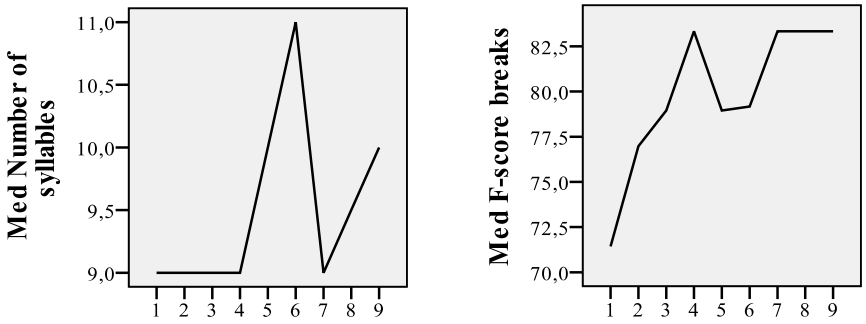
Graph 3. Phoneme perception depending on phonetic environment

Quite naturally, it was found that phonetic environment had affected phoneme perception. Particularly, a strong dependency was found between the phonetic environment and vowels recognition results. It might be interesting to further analyze specific features influencing phoneme perception. However, this task was out of range of our research, as three different sentence sets were created only with the intention to eliminate possible learning effect. Therefore, we didn't thoroughly analyze this relation.

Sentence # in the record

The analysis didn't show any significant relationship between the position certain sentence had in the final recording and perception. This means that respondents haven't tended to perform worse or better regardless of the sample position. However, the graphs below, representing results for different positions in the sequences, demonstrate that medians

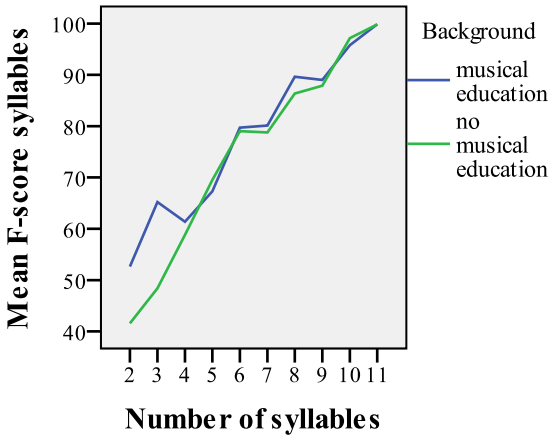
for syllables remembering are higher for positions after 4, with the exception of the sentence #7. Similar picture refers to medians of the F-score for breaks (segmentation task): correctness of recognition consequently gets better as respondents learn (with some drop at positions 5-6 – probably, tired?). This trend justified our choice to shuffle the stimuli within the recordings.



Graph 4. Recall and segmentation depending on the sentence position

4.2.4.3 Intercorrelation between dependent variables

Besides the dependencies revealed between controlled and dependent variables, some relations within dependent variables were also noticed. However, detailed examination showed that almost none of these relations, regardless of their significance, demonstrated any clear dependency.



Graph 5. Correlation between recall amount and accuracy

The only confirmed example was the direct relationship observed between the amount of recall (number of syllables remembered) and recall accuracy (F-score for syllables), see 5. This means that the more syllables were correctly reproduced by respondents, the fewer mistakes had been made, which is quite natural.

Other interrelations did not demonstrate any interesting regularity.

4.2.5 Questionnaires analysis

During the study, not only quantitative data was analyzed, but also some qualitative information based on details given by participants in questionnaires. This information reflected personal attitudes towards different kinds of samples and was not less important and significant in terms of selection of educational settings.

In spite of the quantitative data showing that voice height didn't have any significant effect for the recall or segmentation abilities, respondents found perception of some voice tones more difficult than others. So, 25% of the participants (31% of musicians and 17% of non-musicians) said that the lower voice was most uncomfortable; while for 20% of the participants (19% of musicians and 22% of non-musicians), the most hard was the high pitch; and only 4,5% (1 musician and 1 non-musician) found that the middle voice was the hardest. These results may suggest that, after all, the input stimuli of shifted pitches were somehow distorted, or that these frequencies are, in fact, more difficult for perception.

Evaluating the general impression, 86% of the respondents (85% of musicians and 89% of non-musicians) stated that they found the task of remembering a difficult one. For 22% of the participants (19% of musicians and 28% of non-musicians), the hard part of the test was recognizing the speech flow. In that, 11% of all respondents named both tasks as difficult.

Another important point was connected with the mode of stimuli. So, 59% of the respondents (73% of musicians and 39% of non-musicians) said that the spoken samples were more difficult for perception, while only 9% (4% of musicians and 17% of non-musicians) found that is was harder to perceive the sung stimuli. It is remarkable, that musicians find sung stimuli easier to perceive more often than non-musicians. This argument assumes that musical cues play more important role for people accustomed to music perception. However, the scope of the responders was too small to make any further conclusion.

4.2.6 Analysis summary

One of the most important outcomes of the responses distribution is that respondents, in many cases, were able to successfully write down the source sentences. So, in 42,5% of all cases 11 syllables had been returned (100% recall), while in 61,3% of all cases the output was ≥ 9 syllables (> 80% recall) and in 85,7% cases respondents were able to return ≥ 6 syllables which was more than a half of the input (> 50% recall). And only in 3 cases (0,76% of the

total), which were excluded from the analysis, participants returned empty outputs. In 2 of these cases, the corresponding sentence was the first in the sample consequence and in the third case the sentence's order number was 2. This gives us freedom to explain the bad output by confusion quite natural at the start of an experiment. The said above means that the setting was planned quite well and the challenge of such a difficult task has been overcome.

The analysis showed significant correlations the stimuli mode and the recall and segmentation results, which was, most probably, connected primarily with prosodic cues of the rhythmically organized stimuli rather than with musical features. Voice height, though found perceptionally relevant (as stated by participants), didn't show any significant influence on the dependent variables.

Also, it turned out that phoneme perception depended more on phonetic environment (three sentence sets in our case) than on any other cues. Analysis of the particular phonetic environment features may be very interesting, though, is out of scope of this paper.

5 Conclusion and discussion

Within this study, speech perception in different conditions was examined. The aim of the research was to compare perception results based on the stimuli mode (plain spoken, rhythmically organized spoken or sung stimuli) and pitch (normal, lower and higher) .

The research method was experimental with controlled conditions and a number of variables to be observed. Within the study, an experiment was conducted successfully on 44 participants, who had been asked to listen to 9 recorded sentences in Russian language (unknown to them) and write them down using Latin letters. These 9 sentences were specially prepared using different phonetic environments, voice pitches, rhythmic structures and presentation modes (sung/spoken).

Challenges described in the introduction were mainly overcome, the task was not considered by participants as incredibly difficult, and the results showed that they had coped with the job quite well. Therefore, we were able to collect valuable data, both quantitative (extracted from the responders' outputs and consequently calculated) and qualitative (in the form of questionnaire), and further analysis showed a number of significant results.

To find answers to our research questions, we compared values for various variables and checked them for significance. These variables were introduced to assess variations in the recall / word recognition ability depending on changes in the mode of stimuli presentation and voice height. So, to evaluate the recall results, the measures "amount of recall" and "accuracy of recall" were used; while speech segmentation was assessed using the measure of "segmentation accuracy" based on the word breaks made. The word recognition ability was evaluated by the percentage of correctly recognized phonemes.

To analyse the data, various statistical methods were used. The effect of the voice height on perception couldn't be assessed on per participant basis, so the Kruskal-Wallis and Mann-Whitney tests were used to analyze it. Finally, to evaluate stimuli mode's influence, Friedman's and Wilcoxon's rank tests for related samples were applied.

The analysis showed that the recall ability (for both amount of remembered and recall accuracy), as well as the segmentation accuracy, was affected by the mode of stimuli. Despite our expectation, sung stimuli had been remembered better than spoken ones, even compared to the rhythmically organized condition. The difference between the sung and

rhythmic spoken stimuli (11 syllables vs. 10 syllables, $p < 0.05$) may partly be explained by the duration difference of 2% between the rhythmic spoken and sung samples, while the difference between the plain spoken and rhythmic spoken stimuli (8 syllables vs. 10 syllables, $p < 0.01$) was absolutely obvious and undeniable. This lets us conclude, after some other authors mentioned above, that language prosody cues are more important for speech comprehension than melodic features. However, we cannot indiscriminately put those music features aside; it would be useful to recheck the results on a wider sample.

Concerning speech segmentation, it was found that rhythmic organization definitely made the segmentation task easier ($p < 0.001$), while the difference between the sung and the rhythmic spoken stimuli was not significant. Therefore, answers to the first two research questions of the study are: *yes, mode dependency was found in the stimuli perception results, it had been obviously easier to recall and segment rhythmically organized sentences compared to the plain spoken stimuli; however, benefits of the sung mode were not apparent for segmentation and may be present (to be rechecked) for recall.*

Voice height was found to influence perception of the phonetic image, affecting recognition of vowels in stressed positions and consonants both in stressed and unstressed positions: the lower pitch affected (compared to normal) perception of consonants in stressed positions ($p < 0.05$), while the higher pitch affected (compared to normal pitch) perception of consonants in unstressed position ($p < 0.05$) and (compared to lower and normal) – perception of vowels in stressed position ($p < 0.001$). The results turned out to be much better in the middle tone. Quite surprisingly, a number of tests comparing the effect of the voice height on samples of different modes showed that a tendency of high pitch to affect vowel perception was much stronger for the plain spoken samples, which contradicted with earlier suggestions this effect to be true only for singing (e.g. Sundberg, 1987).

In this concern, we need to point out that the pitch shift was not considerable and in case of more radical change results might have been more distinctive. However, it is possible to assume that, to some extent, the voice height does affect phonetic image perception and, as a consequence, word recognition, as correctly recognized stressed vowel is one of important cues for word distinguishing. Therefore, the said effect of the higher voice should be taken into account when, for example, preparing language learning materials, because ability to correctly recognize phonemes in a word, or its phonetic image, is essential for speech recognition. Thus, the answer to our third research question is: *yes, there was found a clear dependency of phonetic image perception on the stimuli pitch, with the higher pitch being*

considerably worse for recognition, especially in the plain spoken mode; however, the voice tone didn't influence the recall and segmentation abilities.

Although the above conclusions make it apparent that musical cues in such setting do not have a considerable effect for speech recall/word recognition, it doesn't mean that there is no good to use music for educational purposes. Personal attitudes showed by participants in their questionnaires name the sung stimuli as most comfortable for perception. Some specifically stated in the comments that it would be nice to learn languages through music. This emotional reaction is not less meaningful than objective numbers, as learning is a complex process requiring high enthusiasm and motivation.

In this context, it would be interesting to conduct a long-term research investigating the progress of different language groups using in the learning process a lot of songs, poems or plain texts correspondingly. To support using songs in language learning environments, it is also good to say that it is much easier to find songs and their lyrics than any kind of narrated text with a script.

Another area of further research – in the field of linguistics and phonetics – might be analyzing various phonetic environments influencing phoneme perception. Also it might be interesting to try and use in the similar setting other types of melodies and to look at tonal context impact on phoneme perception, or other types of rhythm patterns.

Summarizing the said above, we can conclude that results shown in this study correspond with other studies and make some contribution to the research. Therefore, the aims of the study have been achieved and its research questions have been given answers. The selected research methods have proved to be relevant and corresponding with the task. Data obtained in the study and experimental design developed for it may be used both for further research purposes and for creating educational settings.

References

Ashcraft, M. (2006). Learning and Remembering. *Cognition* , 211-257.

Bahrick, L., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology , 36*, 190-201.

Besson, M., & Schön, D. (2003). Comparison between language and music. *Neurosciences and music* , 269-293.

Bigland, E., Tillmann, B., Poulin, B., Madurell, F., & D'Adamo, D. (2001). The effect of harmonic context on phoneme monitoring in vocal music. *Cognition , 81*, B11-B20.

Bonnel, A., Faita, F., Peretz, I., & Besson, M. (2001). Divided attention between lyrics and tunes of operatic songs: Evidence for independent processing. *Perception & Psychopphysics , 63*, 1201-1213.

Brown, S. (2001). Are music and language homologues? *Annals of the New York Academy of Science , 930*, 372-374.

Brown, S., Martinez, M., & Parsons, L. (2006). Music and language side by side in the brain: a PET study of the generation of melodies and sentences. *European Journal of Neuroscience , 23*, 2791-2803.

Butzlaff, R. (2000). Can music be used to teach reading? *Journal of aesthetic education, 34 (3-4)* , 167-178.

Chazin, S., & Neuschatz, J. S. (1990). Using a mnemonic to aid in the recall of unfamiliar information. *Percept Mot Skills* (71), 1067–71.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America , 24*, 597–606.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences , 37*, 318–325.

Creelman, C. D. (1957). Case of the unkown talker. *Journal of the Acoustical Society of America , 29*, 655.

Crowder, R., Serafine, M., & Repp, B. (1990). Physical interaction and association by contignuity in memory for the words and melodies of songs. *Memory&Cognition* (18(5)), 469-76.

Cutler, A., Norris, D., & Williams, J. N. (1987). A note on the role of phonological expectation in speech segmentation. *Journal of Memory and Language , 26*, 480–487.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America , 27*, 769-773.

Dowd, A., Smith, J., & Wolfe, J. (1998). Learning to Pronounce Vowel Sounds in a Foreign Language using Acoustic Measurements of the Vocal Tract as Feedback in Real Time. *Language and Speech , 41* (1), 1-20.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition , 37* ((1)), 1-9.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language , 16*, 477–501.

Friedrich, C., Kotz, A., Friederici, A., & Alter, K. (2004). Pitch modulates lexical identification in spoken word recognition: ERP and behavioral evidence. *Cognitive Brain Research , 20*, 300-308.

Gfeller, K. (1983). Musical mnemonics as an aid to retention with normal and learning disabled students. *Journal of Music Therapy* , 179-189.

Ginsborg, J., & Sloboda, J. A. (2007). Singers' recall for the words and melody of a new, unaccompanied song. *Psychology of Music , 35* (3), 421-440.

Grimshaw, G., & Yelle, S. (2008). Hemispheric specialization for spoken and sung speech. *Brain and Cognition , 67*.

Gromko, J. (2005). The effect of music instruction on phonemic awareness in beginning readers. *Journal of Research in Music Education, 53 (3)* , 199-209.

Healy, A. F., & Cutting, J. E. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior , 15*, 73–83.

Hollien, H., Mendes-Schwartz, A., & Nielsen, K. (1999, April 13). *Perceptual confusions of high-pitched sung vowels.* Retrieved from www.ScienceDirect.com.

Howie, J., & Delattre, P. (1962). An experimental study of the effect of pitch on the intelligibility of vowels. *The National Association of Teachers of Singing Bulletin , 18* (4), 6-9.

Huron, D., & Ollen, J. (2003). Agogic Contrast in French and English Themes: Further Support for Patel and Daniel. *Music Perception , 21*, 267-272.

Iversen, J. R., Patel, A., & Ohgushi, K. (2006). "How the Mother Tongue Influences the Musical Ear". *4th ASA/ASJ Joint Meeting*, *Popular version of paper 3aPP5.* Honolulu, HI.

Jakobson, R., Fant, C. G., & Halle, M. (1952). *Preliminaries to Speech Analysis.* Cambridge, MA: MIT Press.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review , 96* (3), 459-491.

Jusczyk, P., & Luce, P. (2002). Speech Perception and Spoken Word Recognition: Past and Present. *Ear & Hearing , 23* (1), 2-40.

Kilgour, A., Jakobson, L., & Cuddy, L. (2000). Music training and rate of presentation as mediators of text and song recall. *Memory & Cognition , 28* (5), 700-710.

Koelsch, S., & Siebel, W. (2005). Towards a neural basis of music perception. *Trends in Cognitive Science , 9*, 578-584.

Koelsch, S., Gunter, T., Wittfoth, M., & Sammler, D. (2005). Interaction between Syntax Processing in Language and Music: An ERP Study. *Journal of Cognitive Neuroscience , 17* (10).

Kolinsky, R., Lidji, P., Peretz, I., Besson, M., & Morais, J. (2009). Processing interactions between phonology and melody: vovels sing but consonant speak. *Cognition 112* , 1-20.

Kouri, T., & Telander, K. (2008). Children's reading comprehension and narrative recall in sung and spoken story context. *Child Language Teaching and Therapy* .

Kouri, T., & Winn, J. (2006). Lexical Learning in Sung and Spoken Story Script Context. *Child Language Teaching and Therapy , 22* (3), 293-313.

Krashen, S., & Terrell, T. (1983). The natural approach: language acquisition in the classroom. *Alemany Press* .

Kuhl, P. K., Andruski, J. E., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., et al. (1997). Cross-language analysis of phonetic units addressed to infants. *Science , 277*, 684–686.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of Acoustical Society of America , 29*, 98–104.

Lattner, S., Meyer, M., & Friederici, A. (2005). Voice Perception: Sex, Pitch, and the Right Hemisphere. *Human Brain Mapping , 24* (1), 11-20.

Liberman, A. M., DeLattre, P. D., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *American Journal of Psychology , 65*, 497–516.

Liberman, A. M., Harris, K. S., Eimas, P. D., Lisker, L., & Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonetic significance. *Language and Speech , 54*, 175–195.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology , 54*, 358–368.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. L. (1961). The discrimination of relative-onset time of the components of certain speech and non-speech patterns. *Journal of Experimental Psychology , 61*, 379–388.

Maess, B., Koelsch, S., Gynter, T., & Frederici, A. (2001). Musical syntax is processed in Broca's area: An MEG study. *Nature Neuroscience , 4*, 540-545.

Magne, C., Aramaki, M., Astesano, C., Gordon, R., Ystad, S., Farner, S., et al. (2004). Comparison of rhythmic processing in language and music: An interdisciplinary approach. *Journal of Music and Meaning , 3*.

Maidhof, C., & Koelsch, S. (2010). Efects of Selective Attention on Syntax Processing in Music and Language. *Journal of Cognitive Science* , 1-16.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review , 79*, 124–145.

McElhinney, M., & Annett, J. M. (1996). Pattern of efficacy of a musical mnemonic on recall of familiar words over several presentations. *Perceptual and Motor Skills* (82(2)), 395-400.

McNeill, D., & Lindig, K. (1973). The perceptual reality of the phoneme, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior , 12*, 419–430.

Medeiros, B. (2008). Intonational Aspects of Songs and Song Competence. *Genebra: EMUS organização* .

Medina, S. L. (1990). The effects of music upon second language vocabulary acquisition. *Annual Meeting of the Teachers of English to speakers of other languages.* San Francisco.

Mills, C. B. (1980). Effects of context on reaction time to phonemes. *Journal of Verbal Learning and Verbal Behavior , 19*, 75–83.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception & Psychophysics , 18*, 331–340.

Morozov, V. (1965). Intelligibility in singing as a function of fundamental voice pitch. *Soviet Physics Acoustics , 10*, 279-283.

Nakada, T., & Abe, J. (2005). Text-melody asymmetrical integration in memory for songs: Contributions of rhythm and pitch patterns of melodies. *Tech.Rep.* (35), 1-47.

Nettl, B. (2000). An ethnomusicologist contemplates universals in musical sound and musical culture. *The origin of music* , 463-472.

Palmer, C., & Hutchins, S. (2006). What is musical prosody? *Psychology of Learning and Motivation , 46*, 245-278.

Patel, A. (2006). An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies: A Comment on J.G.S. Pearl's “Eavesdropping with a Master: Leos Janácek and the Music of Speech.”. *Empirical Musicology Review , 1* (3), 166-169.

Patel, A., & Daniele, J. (2002). An empirical comparison of rhythm in language and music. *Cognition , 87*, B35-B45.

Pitt, M., & Samuel, A. (1990). Attentional Allocation during Speech Perception: How fine is focus? *Journal of Memory and Language , 29*, 611-632.

Pollack, I. (1952). The information in elementary auditory displays. *Journal of the Acoustical Society of America , 24*, 745– 749.

Poulin-Charronnat, B., Bigand, E., Madurell, F., & Peereman, R. (2005). Musical structure modulates semantic priming in vocal music. *Cognition , 94* (3), B67-78.

Quen'e, H., & Port, R. F. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica 62 (1),* , 1-13.

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition , 35 (2)*, 242-253.

Rainey, W. D., & Larsen, J. D. (2002). The Effect of Familiar Melodies on Initial Learning and Long-term Memory for Unconnected Text. *Music Perception , Vol. 20, No. 2*, 173–186.

Ross, D., Choi, J., & Purves, D. (2007, April 5). Musical intervals in speech. *Center for Cognitive Neuroscience and Department of Neurobiology* . Durham: Duke University.

Rubin, D. (1995). *Memory in oral traditions: The cognitive psychology of counting-out rhymes, ballads, and epics.* New York: Oxford University Press.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science , 274*, 1926–1928.

Saffran, J., Johnson, R. E., Aslin, N., & Newport, E. L. (1999). Abstract Statistical learning of tone sequences by human infants and adults. *Cognition , 70*, 27–52.

Salcedo, C. (2010). The Effects of Songs in the Foreign Language Classrum On Text Recall, Delaed Text Recall and Involuntary Mental Rehearsal. *Journal Of College Teaching & Learning (TLC) , 7* ((6)).

Samson, S., & Zatorre, J. (1991). Recognition memory for text and melody of songs after unilaterial temporal lobe lesion: Evidence for dual encoding. *Journal of Experimental Psychology: Learning,memory, and cognition , 17*, 793-804.

Sanders, L., & Neville, H. (2000). Lexical, syntactic, and stress-pattern cues for speech segmentation. *Journal of Speech, Language, and Hearing Research: JSLHR, 43 (6)* , 1301-21.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior , 9*, 295–302.

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition , 106*, 975-983.

Schön, D., Gordon, R., Campagne, A., Magne, C., Astesano, C., Anton, J., et al. (2010). Similar celebral networks in language, music and song perception. *NeuroImage , 51* (1), 450-461.

Schön, D., Leigh Gordon, R., & Besson, M. (2005). Musical and linguistic processing in song perception. *The Neurosciences and Music II: From Perception to Performance* , 71-81.

Schön, D., Magne, C., & Besson, M. (2004). The music of speech:Music facilitates pitch processing in language. *Psychophysiology , 41*, 341-349.

Scotto di Carlo, N., & Germain, A. (1985). A perceptual stady of the influence ofpitch on the intelligibility of the sung vowels. *Phonetica , 42*, 188-197.

Selkirk, E. (1984). *Phonology and syntax: the relation between sound and structure.* Cambridge: Mass. : MIT Press.

Serafine, M., Crowder, R., & Repp, B. (1984). Integration of melody and text in memory for songs. *Cognition , 16*, 285-303.

Serafine, M., Davidson, J., Crowder, R., & Repp, B. (1986). On the nature of melody-text integration in memory for songs. *Journal of Memory and Language , 25*, 123-135.

Smith, L., & Scott, B. (1980). Increasing the intellegibility of sung vowels. *Journal of the Acoustical Society of America , 67*, 1795-1797.

Sternberg, R. J., & Mio, J. (2008). *Cognitive psychology.* Cengage Learning.

Stevens, K., & Keller, P. (2001). Discriminating Pitch Contour in Words and Music: A Comparison of Thai and English Speakers. *Australian Journal of Psychology , 53*, Supplement.

Stumpf, C. (1926). Die Sprachblaute. *Berlin and New York:'Springer-Verlag* .

Sundberg, J. (1970). Formant structure and articulation of spoken and sung vowels. *Folia Phoniatrica , 12*, 28-48.

Sundberg, J. (1987). How pitch affects phoneme recognition. *The Science of the Singing Voice* .

Swinney, D. A., & Prather, P. (1980). Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts. *Perception & Psychophysics , 27*, 104–110.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America , 60*, 198–212.

Wallace, W. (1994). Memory for Music: Effect of Melody on Recall of Text. *Journal of Experimental Psychology: Learning, Memory, and Cognition , Vol. 20, No. 6*, M71-M85.

Wigram, T., & Gold, C. (2006). Music therapy in the assessment and treatment of autistic spectrum disorder: clinical application and research evidence. *Child: Care, Health and Development , 32* (5), 535-542.

Wolfe, J. (2002). Speech and music, acoustic and coding, and what music might be 'for'. *7th International Conference on Music Perception and Cognition, ,.* Sydney.

Ystad, S., Magne, C., Farner, S., Pallone, G., Aramaki, M., Besson, M., et al. (2007). Electrophysiological Study Of Algorithmically Processed Metric/Rhythmic Varistions in Language and Music. *Journal on Audio, Speech, and Music Processing* .

Zheng, X., & Pierrehumert, J. (2010). The effects of prosodic prominence and serial position on duration perception. *The Journal of the Acoustical Society of America* .

Appendix A. Samples

Sentence set #1	Sentence set #2	Sentence set #3
1.1 Dáma tóli súku díka padalá	2.1 Sála dóma nétu típa malakó	3.1 Mála tími soku dúri panisú
1.2 Dáli tóni súpu dímu palilá	2.2 Sáni dóli nétu tína dalikó	3.2 Máni tína kóni dúhi kalisú
1.3 Dána padalá tómi kasú díka	2.3 Sámi dalikó dóma putí nétu	3.3 Mámi palasú kóli vidú tíha

Sample sequences recorded as tracks to be played during the experiment

The letter “h” denotes high-tone samples, “m” – middle tone, “l” – low tone.

#1 **1.1h** – 2.2m – 3.3l – 1.2h – 2.3m – 3.1l – 1.3h – 2.1m – 3.2l

#2 **1.1m** – 2.2l – 3.3h – 1.2m – 2.3l – 3.1h – 1.3m – 2.1l – 3.2h

#3 **1.1l** – 2.2h – 3.3m – 1.2l – 2.3h – 3.1m – 1.3l – 2.1h – 3.2m

#4 2.2m – 3.3l – 1.2h – 2.3m – 3.1l – 1.3h – 2.1m – 3.2l – **1.1h**

#5 2.2l – 3.3h – 1.2m – 2.3l – 3.1h – 1.3m – 2.1l – 3.2h – **1.1m**

#6 2.2h – 3.3m – 1.2l – 2.3h – 3.1m – 1.3l – 2.1h – 3.2m – **1.1l**

#7 3.3l – 1.2h – 2.3m – 3.1l – 1.3h – 2.1m – 3.2l – **1.1h** – 2.2m

#8 3.3h – 1.2m – 2.3l – 3.1h – 1.3m – 2.1l – 3.2h – **1.1m** – 2.2l

#9 3.3m – 1.2l – 2.3h – 3.1m – 1.3l – 2.1h – 3.2m – **1.1l** – 2.2h

#10 1.2h – 2.3m – 3.1l – 1.3h – 2.1m – 3.2l – **1.1h** – 2.2m – 3.3l

#11 1.2m – 2.3l – 3.1h – 1.3m – 2.1l – 3.2h – **1.1m** – 2.2l – 3.3h

#12 1.2l – 2.3h – 3.1m – 1.3l – 2.1h – 3.2m – **1.1l** – 2.2h – 3.3m

#13 2.3m – 3.1l – 1.3h – 2.1m – 3.2l – **1.1h** – 2.2m – 3.3l – 1.2h

#14 2.3l – 3.1h – 1.3m – 2.1l – 3.2h – **1.1m** – 2.2l – 3.3h – 1.2m

#15 2.3h – 3.1m – 1.3l – 2.1h – 3.2m – **1.1l** – 2.2h – 3.3m – 1.2l

#16 3.1l – 1.3h – 2.1m – 3.2l – **1.1h** – 2.2m – 3.3l – 1.2h – 2.3m

#17 3.1h – 1.3m – 2.1l – 3.2h – **1.1m** – 2.2l – 3.3h – 1.2m – 2.3l

#18 3.1m – 1.3l – 2.1h – 3.2m – **1.1l** – 2.2h – 3.3m – 1.2l – 2.3h

#19 1.3h – 2.1m – 3.2l – **1.1h** – 2.2m – 3.3l – 1.2h – 2.3m – 3.1l

#20 1.3m – 2.1l – 3.2h – **1.1m** – 2.2l – 3.3h – 1.2m – 2.3l – 3.1h

#21 1.3l – 2.1h – 3.2m – **1.1l** – 2.2h – 3.3m – 1.2l – 2.3h – 3.1m

Appendix B. Questionnaire

Listening experiment **Track number _____**

Instructions

Please, listen to the sound file which consists of 9 short sentences (spoken or sung).
They are composed like follows:

Beep – **Sent1** – *7 sec. pause* – **Sent1** – *15 sec. pause* – Beep – **Sent2** – *7 sec. pause* – **Sent2** – etc.

The task is to write down what you hear. Do not try to understand the sentences, they are in a language you don't know.
Just **write down word by word** what you hear, **in Latin letters** (a, d, k..., like "galava", “kura"...).
Do not think about the letters choice, just try to fix every word exactly **as you hear it**, not thinking about what should be right. "Right" is what you perceive.

1.

_____

2.

_____

3.

_____

4.

_____

5.

_____

6.

_____

7.

_____

8.

_____

9.

_____

Please, turn the page

Please, fill in the following form:

Age _____ **Gender** female □ male □

What do you study? Level? _____

Do you have musical education? **Yes** □ **No** □

If yes, what level? _____

What languages do you speak? _____

Was there anything difficult to you? What? _____

_____

Did you find some of the sentences more difficult to perceive than others?

Yes □ **No** □ **If yes, which kinds were more difficult?** □ Spoken

□ Sung

□ Low voice

□ Middle voice

□ High voice

Were the difficulties connected with □ **remembering** **or** □ **recognizing?**

Any notes, comments, suggestions? _____

_____

_____

THANK YOU VERY MUCH

for your contribution! It is very much appreciated!! Good Luck!