Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Speech and Music Discrimination: Human Detection of Differences between Music and Speech based on Rhythm

Madeleine Stanev, Johannes Redlich, Christian Knörzer, Ninett Rosenfeld, Athanasios Lykartsis

Audio Communication Group, Technische Universität Berlin, Germany

madeleine.stanev@mailbox.tu-berlin.de, johannes.redlich@mailbox.tu-berlin.de,
christian.knoerzer@mailbox.tu-berlin.de, ronin@mailbox.tu-berlin.de,
athanasios.lykartsis@tu-berlin.de

Abstract

Rhythm in speech and singing forms one of its basic acoustic components. Therefore, it is interesting to investigate the capability of subjects to distinguish between speech and singing when only the rhythm remains as an acoustic cue. For this study we developed a method to eliminate all linguistic components but rhythm from the speech and singing signals. The study was conducted online and participants could listen to the stimuli via loudspeakers or headphones. The analysis of the survey shows that people are able to significantly discriminate between speech and singing after they have been altered. Furthermore, our results reveal specific features, which supported participants in their decision, such as differences in regularity and tempo between singing and speech samples. The hypothesis that music trained people perform more successfully on the task was not proved. The results of the study are important for the understanding of the structure of and differences between speech and singing, for the use in further studies and for future application in the field of speech recognition.

Index Terms: speech-music discrimination, speech perception, speech rhythm, computational paralinguistics

1. Introduction

The identification of rhythm in language as it can be detected in music has been a growing field in many disciplines such as linguistics, music psychology or neuroscience. The nature and understanding of rhythm is vital in order to locate similar patterns in speech. Even though periodicity has been widely linked to the description of rhythm, it is important to treat this term carefully, especially with regard to speech rhythm [1]. Also according to Patel [1], rhythm can be best described as "the systematic pattern of sound in terms of timing, accent and grouping" since "both speech and music are characterized by systematic temporal, accentual, and phrasal patterning". Rhythm in speech, on the other hand, refers to the way languages are organized in time. Most linguists argue that certain languages differ in their rhythmic structure. The basis of the exact distinction is yet hard to define. It is assumed that the differences depend on whether sequences of syllables or stresses are isochronous. English and Dutch, for example, manifest highly varied syllable patterns and therefore are referred to as "stress-timed", whereas languages like French and Spanish are "syllable-timed" due to their low varied syllable structure [2]. Nevertheless, critics argue that these distinctions are nonexistent [3]. Lehiste [4], for example, mentioned that isochrony is considered as a perceptive phenomenon. In recent years, studies have turned away from the idea of isochrony and, to a greater extent, have focused on vocalic and intervocalic intervals in speech [3, 5, 6]. This shift away from classes and towards rhythmic differences has evoked the idea of a rhythmic continuum in which the clusters become blurry if more languages are added [2, 7].

The identification of durational vowel patterns in speech requires the application of specific methods. The development of the *Pairwise Variability Index (nPVI)* by Patel & Daniele [8] allowed the researchers to quantify prosody, which was directly comparable to music. With the nPVI the durational contrast between successive elements in a sequence can be detected and measured. The method has been widely applied to exploit more accurate differences between "stress-timed" and "syllable-timed" languages. Vowel-based measures such as the nPVI are the most plausible way to compare speech to music since musical notes can be roughly compared to syllables, with vowels forming the core of syllables [9].

Ramus and Mehler [10] examined the ability of infants that have been raised in a bi- or multilingual environment to discriminate between languages. They used *speech resynthesis* to further explore the components of prosody by measuring all relevant acoustic parts of a speech signal and subsequently resynthesizing the speech material with an appropriate algorithm. This enabled them to freely select or dispose of the components they wanted to use, such as phonemes, rhythm, and intonation. The results showed that when preserving intonation, rhythm, and broad phonetic categories alternately, that the use of these suprasegmental cues is indeed a sufficient way to allow the discrimination of languages.

Ohala and Gilbert [11] tried a different approach by providing participants with conversations instead of single sentences. The use of long passages instead of short ones had no effect on participants' performances. They furthermore converted the speech signal with the help of a voltage-controlled-frequency and voltage-controlled-amplitude signal generator into a 'buzz' while still maintaining the same frequency, timing, and amplitude of the original signal. The results were highly significant statistically ($p < 0.0001$) but the scores were still low enough to indicate that the signal lacks crucial prosodic information after the conversion. Syllables and word boundaries, for example, are not recognizable anymore. Since syllables may be the closest comparable cue to a music tone due to their salience for perception, their preservation might lead to a better performance.

10.21437/SpeechProsody.2016-46

Several studies [11, 12, 13, 14] have succeeded in demonstrating that the rhythmic pattern in speech and language is able to act as clue even after the extraction of different prosodic components. In this paper we aim to further investigate this subject, also with regard to the rhythmic structure to music. Instead of solely comparing music with spoken sentences, we make use of singing sequences without any instrumental background (a capella singing). Thus, the signals are quite similar with regard to their prosodic structure and are barely left with any musical features. This gives us the opportunity to explore the rhythmic patterns more specifically without the distraction of other cues or the dominance of instrumental rhythm. However, discoveries have been already made regarding the phenomenon of musicians composing their music on the basis of language-specific-rhythmic patterns [15]. By overlaying the envelope curve of the speech and singing sequences with a carrier signal we are able to eliminate any other linguistic information apart from rhythm, allowing us to further explore rhythmical characteristics and differences of speech and music.

2. Method

The question is: "Are people able to distinguish singing from spoken words only by the rhythm?". Based on personal experience, the current status of research and the experiments of Ramus et al. [14], we find the investigation of this question worthwhile. According to Ramus et al., speech has a significant time structure. We will use this time structure to test if people are able to distinguish singing from spoken words. Therefore, the *H0* hypothesis is that singing and speech are distinguishable based only on rhythmic cues, whereas *H1* states that it is not possible to distinguish them better than chance on a significant level. To that purpose we used recordings from a multilingual prosodic corpus (MULTEXT PD) [16]. The corpus derived from the EUROM 1 speech database that was developed within the Esprit SAM (Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation) project [17]. The collection contains 40 different thematically connected passages, each comprising five sentences. 50 different speakers read in five languages (French, English, Italian, German and Spanish). We used a mere 2 out of the 40 passages, spoken by 10 different speakers. Each speaker spoke one of these two passages so that we got 10 passages. With this selection we want to get a pool of different ways of speaking, to represent different rhythms of speaking. Furthermore, we only made use of the German passages since most of our participants were of German nationality, so that we are primarily researching rhythm perception for native speakers. The acoustic quality of the recording is high with sampling rate at 20 kHz, 16 bits, and the recordings having been performed in an anechoic room. An example of a recorded message can be seen below:

> I have a problem with my water softener. The water level is too high, and the overflow keeps dripping. Could you arrange to send an engineer on Tuesday morning, please? It's the only day I can manage this week. I'd be grateful if you could confirm the arrangement in writing.

We downloaded the singing files from youtube. Any videos uploaded on youtube are publicly available and can therefore be legally downloaded and processed for research purposes. Furthermore, the files are altered in an unrecognizable way. We used five a capella songs, three choir sequences, and two rap songs. They were all in German language in order to ensure a better comparison basis with the spoken sentences.

To create the stimuli we followed a self-developed method, because of transparency and simplicity in the algorithm and the accuracy of the envelope. It is also possible to produce the envelope by using the Hilbert transform, but in our case it was not so well fitted to the signal. We created the stimuli by overlaying a carrier signal (noise or sine tone) with the envelope curve of the speech and singing signals. The noise is in our case a band-pass filtered noise between 200 Hz and 400 Hz. Figure 1 shows the detection of the peak points. The detection by itself was accomplished with a MATLAB script. Connecting the peak points of the useful signal formed the envelope curve of a signal. The distance between the peak points was defined by the desired maximum frequency of 2000 Hz of the envelope curve.

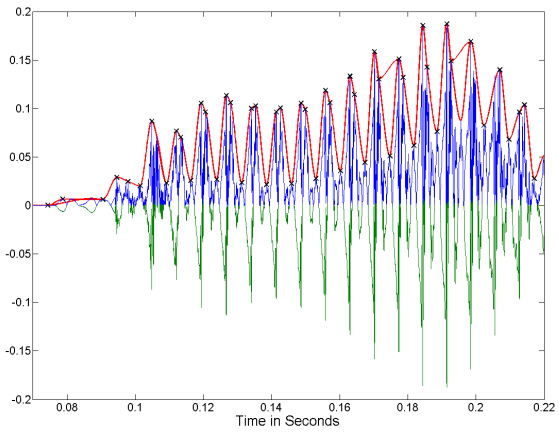

Figure 1: *Generation of the envelope curve for a stereo signal. The green curve is the speech-signal, the blue curve is the absolute value of the speech-signal, black crosses represent signal peaks and the red curve is the interpolated envelope curve.*

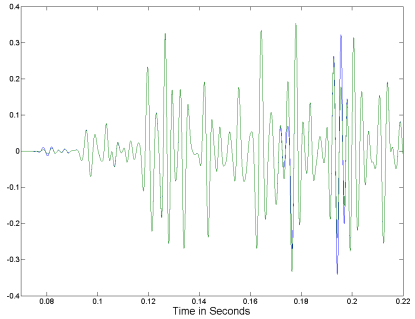

Figure 2: *Final stimuli with sine wave as carrier.*

The connection of the points was generated with a piece-by-piece cubic interpolation creating a high frequency useful signal on a low frequency carrier signal. This method can have as a result the appearance of unintended artifacts, like "bubbling" or "scratching" sounds, which we will accept due to its otherwise high effect of eliminating every linguistic aspect but the rhythm and the outcome of a reduced signal.

The results did not show any evidence of participants feeling disturbed by these artifacts. We are convinced that other methods could bring fewer artifacts, but then the extraction of the rhythm could be less effective. During the survey participants are asked to comment on their perception of the stimuli, to gain a deeper understanding of the applicability. We decided to limit the lengths of the stimuli to ten seconds, which still ensured the detectability of the rhythm by simultaneously preserving the participants from overstimulation or fatigue. We decided to conduct our research via the online survey platform LimeSurvey to reach more potential participants. A recent study by Pysiewicz [18] suggested that results produced in online settings hold as much validity as the ones within laboratory settings. Since this research was conducted in the field of auditory perception research, it can be applied here. The survey was also designed to query about further personal information such as the degree of musical education or the valuation of musical understanding. This information may help to further specify any influential aspects on the ability to discriminate. Since the study was distributed via the Internet, the participants could choose their own test setting, which gave us no influence on the environment. They used their own equipment, either loudspeakers or headphones, to listen to the sound samples. This is not a disadvantage, since the perception of the rhythm as a high-level element is relatively disconnected from the quality of reproduction. However, a test phase with two stimuli, pure sine tone and pure noise, allowed the participants to adjust the volume accordingly and to get a general idea of the transformed sequences. In the next step, 20 out of the 40 possible testing samples were presented randomly with regard to the alteration pattern (sine or noise), alternating between spoken sentences and singing. The randomization of the stimuli counteracted the appearance of any potential systematic sequential effects or response biases. To get a better understanding of the clues that had helped participants with their decision, they were asked for further explanations in the end of the survey. Thirty-two volunteers participated in the survey. 84.4% of the contestants completed the survey in German and 15.6 % in English. 78.1 % were German native speakers and the remaining 21.9 % were almost equally distributed between Spanish, Russian, Chinese, Japanese and Greek. Participants were mostly male (65.6%) than female (34.4%), with a mean age of 28 years.

3. Results

Table 1 presents a first overview of our results including hit rates, false alarm rates, discrimination scores A' [19] and the response bias measures B''_D [20]. Hit rates serve as the percentage of correctly identified singing sequences and false alarm rates include the percentage of speech passages that were selected as singing.

Table 1. *Hit rates, false alarm rates, discrimination scores, and response bias measures for each stimuli condition. A' was compared to 0.5 (chance level).*

Signal Carrier	Hit rates	False Alarms	A'	B''_D
Sine	0.78	0.27	0.84	-0.135
Noise	0.66	0.18	0.83	0.402

To perform an analysis of discrimination, we first applied a Kolmogorov-Smirnov test for normality to the discrimination scores. The results showed that the distribution of the score could be considered as normal with all p values < 0.134. Afterwards, a t-Test was computed to compare the discrimination scores to the chance level of 0.5. The results showed that the discrimination scores were significantly above chance level in the sine ($t(16) = 4.39$, $p \leq 0.001$) and noise ($t(16) = 4.523$, $p \leq 0.001$) conditions. Hence, participants were able to distinguish between singing and speech regardless of its transformation, confirming the *H0*. According to the results, we assume that their decisions were not random but based on the detection of meaningful cues.

To gain a deeper knowledge on any potential effects of music education on the ability to discriminate, we performed an ANOVA. Test results show with $p>0.462$ no significance. Therefore, music education or training does not correlate with a better performance on the test. Results with the regard to age, sex and nationality showed no significance either. After the discrimination task we provided participants with five possible cues that could have helped with their decision and asked them to determine the ones that had contributed to their choice. Most participants opted for the sound-pause-ratio (81.25%) followed by regularity (62.5%), accentuation (56.25%), timing (46.9%), and grouping (18.75%). When asked to further define the aspects that lead participants to a certain conclusion, several stated that they focused on regularity or periodic sequences. Irregular patterns and a faster tempo were assumed to belong to speech passages, whereas more regular patterns and the impression of fluency were associated with singing sequences. A few participants referred to perceived rhythm as a main characteristic.

4. Discussion

The results presented in Table 1 indicate that participants were able to discriminate between singing and speech in both conditions. Ramus et. al [5] noted that for single frequencies, intonation-stressed pitches and syllables were still detectable. In addition, signals without any intonation cues showed that "syllabic rhythm was a robust cue for discrimination" [5]. The B''_D scores show a slight difference of response bias between the two conditions. Participants were conservative with the noise stimuli and liberal with the sine stimuli, meaning that the tendency to recognize the noise stimuli was slightly higher with the sine stimuli. This could relate to the clearer acoustic nature of the sine stimuli in comparison to the noise ones. The test results of the ANOVA implied that music education in any form has ultimately no impact on the discrimination of the stimuli. The right perception of stimuli seems to be independent from musical preparatory training or talent. A reason for this could be the different significance of rhythm in speech and singing. Human beings each learn at least a language in their lifetime and are native speakers of it. This circumstance allows them to develop a special consciousness for speech rhythm. According to Patel [1], speech rhythm is determined by its irregularity and musical patterns are defined by its periodicity. Therefore, musical patterns seem to be structured in a more elementary way than speech patterns. We suppose that simpler patterns are in general also easier to detect. Assuming that people have the same knowledge with regard to complex speech patterns, no difference exists between musically trained people and non-musicians. Hence, participants mostly relied on the sound-pause-ratio while listening to the stimuli. Since rhythm correlates with temporal

aspects, the explicit focus on that ratio seems to be meaningful. Regularity plays another important role and is often associated with rhythmic structures in music. Recurring periodic patterns served as a hint to the existence of some musical structure.

5. Conclusions

The aim of this paper was to test the capacity of subjects to discriminate between speech and singing with rhythm serving as the only indicator. While most studies have been focusing on the discrimination of different languages, especially with infants [21, 22], or the connection between language and corresponding rhythms in compositions [13], researchers are still at odds with the exact definition of rhythm in speech. Therefore, we examined the musical rhythmic structure as well as the linguistic ones. Even though our results resemble the previous ones, we could discover that rhythmic structures vary between language and music, as participants stated these variations as a crucial criterion for the discrimination of presented signals. As mentioned above, participants rated signals regarding to their regularity and periodicity. Participants found irregularity in rhythmic patterns and faster pace to be indicators for speech whereas they characterized singing sequences by regular patterns and a medium pace. The rhythm of speech is defined by the use of syllables and stresses depending on the context, systemizing the sequences of sounds in the sense of time, accentuation and grouping [1]. As a consequence, language is never just a periodic repetition of linguistic units [1]. In contrast, Western music has a regular rhythmic structure, which allows a high complexity on other musical and more important dimensions [1]. Flexibility of pace is often used as a way of expression [1]. We also offer a new method that allows the complete transformation of a signal without losing desired valuable components. By overlaying the signals' envelope curves with sine and noise carrier signals we ensured that all linguistic cues were eliminated besides rhythm. The discoveries in this study are applicable to various fields. An explicit attribution of singing stimuli as such may be, for example, useful for algorithms in applications that aim to match a singing sequence to the appropriate song. We suggest further research to explicitly identify the specific rhythmic patterns that contribute to a correct perception of singing and speech.

A better comprehension of rhythm in speech and singing can be meaningful in the development of speech therapies. Patients' entry into the therapy could be simplified by using the periodic rhythm patterns of singing to give them a better understanding of the rhythmic nature of language. A specific linguistic rhythm can be better suited for a specific, more efficient use of the language in music [23]. It is probably no coincidence that most operatic arias were composed in Italian. However, this assumption contradicts with the categorization of languages into "stress-timed" and "syllable-timed" since the English language is commonly regarded as "most singable". Future studies could draw a comparison between several different languages to further exploit our results. The high performance in our study could derive from specific characteristics of the German language that may simplify discrimination. One should also take into account that we used different singing styles (rap, ballad, etc.). Some of these are more similar to speech (like rap) than others. Therefore, they are not clearly recognizable as singing sequences because the rhythmical features are less typical for musical rhythm structures. It would be interesting to investigate if a more homogenous selection or a selection with a larger variance, consisting of greater number of sequences than used here might offer even clearer results. Furthermore, it would be interesting to see if the test subjects perform differently when only comparing rap (or other more specific vocal singing types) sequences to speech, in contrast to the comparison of a capella singing, so as to determine if the detection of rhythm differences is dependent on the kind of vocal song involved. Likewise, this could also be performed for spontaneous speech rather than read as in our study. Finally, future work could also explore the significance of our results when applied to training with stutterers, as they are often more capable in singing, which could point to a better understanding of rhythm in singing than normal speech.

6. References

[1] Patel, Aniruddh D.: "Rhythm." In: Anriuddh D. Patel (Ed.) *Music, Language and the Brain*. Oxford, The Oxford University Press. pp. 96 – 177, 2007.

[2] Grabe, Esther, and Ee Ling Low: "Durational variability in speech and the rhythm class hypothesis." In: *Papers in laboratory phonology*, 7, pp. 515-546, 2002.

[3] Roach, P.: "On the distinction between 'stress-timed' and 'syllable-timed' languages," In: D. Crystal (Ed.) *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F.R. Palmer*, Edward Arnold, London, pp. 73–79, 1982.

[4] Lehiste, Ilse: "Isochrony reconsidered." In: *Journal of phonetics*, 5 (3), pp. 253 – 263.

[5] Low, Ee Ling, Grabe, Esther, and Francis Nolan: "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English." In: *Language and speech*, 43 (4), pp. 377-401, 2000.

[6] Ramus, Franck, Nespor, Marina, and Jacques Mehler: "Correlates of linguistic rhythm in the speech signal." In: *Cognition*, 73 (3), pp. 265-292, 1999.

[7] Farinas, Jérôme, and François Pellegrino: "Automatic rhythm modeling for language identification." In: *INTERSPEECH*, pp. 2539–2542, 2001.

[8] Patel, Aniruddh D., and Joseph R. Daniele: "An empirical comparison of rhythm in language and music." In: *Cognition*, 87 (1), pp. 35-45, 2003.

[9] Patel, Aniruddh D.: „Rhythm in Language and Music – Parallels and Differences." In: *Annals of the New York Academy of Science*, 999 (1), pp. 140-143, 2003.

[10] Ramus, Franck, and Jacques Mehler: "Language identification with suprasegmental cues: A study based on speech resynthesis." In: *The Journal of the Acoustical Society of America*, 105 (1), pp. 512-521, 1999.

[11] Ohala, J. J., and J. B. Gilbert: "Listeners' ability to identify languages by their prosody." In: *Report of the Phonology Laboratory Berkeley*, 2, pp. 126-132, 1978.

[12] Patel, Aniruddh D., Peretz, Isabelle, Tramo, Mark, and Raymonde Labreque: „Processing Prosodic and Musical Patterns: A Neuropsychological Investigation." In: *Brain and Language*, 61, pp. 123-144, 1998.

[13] Patel, Aniruddh D., Iversen, John R., and Jason C. Rosenberg: „Comparing the rhythm and melody in speech and music: The case of British English and French." In: *The Journal of Acoustical Society of America*, 119 (5), pp. 3034-3047, 2006.

[14] Ramus, Franck, Dupoux, Emmanuele, Zangl, Renate, and Jacques Mehler: „An empirical study of the perception of language rhythm.", 2000.

[15] Hannon, Erin E.: "Perceiving speech rhythm in music: Listeners classify instrumental songs according to language of origin" In: *Cognition*, 111, pp. 403-409, 2009.

[16] Campione, Estelle, and Jean Véronis: "A Multilingual Prosodic Database." In: *ICSLP*, 98, pp. 3163-3166, 1998.

[17] Chan, D., Fourcin, A., Gibbon, D., Grandström, B., Huckvale, M, Kokkinakis, G., Kvale, K., Lamel, L., lindberg, B., Moreno,

A., Mouropoulos, J., Senia, F., Transcoso, I., Velt, C. and Zeiliger, J.: "EUROM – A Spoken Language Ressource for the EU." In: *Proceedings of Eurospeech'95,* Madrid, 1995.

[18] Pysewicz, Andreas: *Validity and Reliability of Internet-based Audiotory Perception Experiments*. MA-Arb. Berlin: Technische Universität, 2014.

[19] Snodgrass, J. G., G. Levy-Berger, and M. Haydon: *Human Experimental Psychology*. New York: Oxford University Press, 1985.

[20] Donaldson, W.: „Accuracy of *d ′* and *A′* as estimates of sensitivity.“ In: *Bulletin of the Psychonomic Society*, 31, pp.271-274, 1993.

[21] Nazzi, Thierry, Josiane Bertoncini, and Jacques Mehler: "Language discrimination by newborns: toward an understanding of the role of rhythm." In: *Journal of Experimental Psychology: Human perception and performance*, 24 (3), pp. 756-766, 1998.

[22] Ramus, Franck: "Language discrimination by newborns: Teasing apart phontactic, rhythmic, and intonational cues." In: *Annual Review of Language Acquisition*, 2, pp. 85-115, 2002.

[23] Schafroth, Elmar: "Sprache und Musik. Zur Analyse gesungener Sprache anhand von Opernarien." (Erweiterte Fassung [S. 33] In: http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=23678.