Rhythm classes and speech perception

Amalia Arvaniti, Tristie Ross

Department of Linguistics, University of California, San Diego, USA

amalia@ling.ucsd.edu, tross@ling.ucsd.edu

Abstract

This study indirectly tests whether American, Greek and Korean listeners can classify low-pass filtered utterances of English, German, Greek, Italian, Korean and Spanish into rhythm classes, by examining how they rate each utterance's rhythm in comparison to a series of non-speech trochees. Such classification was difficult for all groups of listeners and did not support the rhythmic classification of the languages of the stimuli, casting doubt on the impressionistic basis of the rhythm class hypothesis.

Index Terms: speech perception, rhythm class, rhythm

1. Introduction

A classic view of speech rhythm advocates that languages are divided into rhythmic classes, namely stress-, syllable- and mora-timing. This typology has been investigated to a great extent but results so far have failed to strongly support it. Production studies dating from the 1960s to the 1980s focused on the search for isochrony, the (near)-equal duration of the units of each rhythm class but failed to find evidence for it (see Bertinetto 1989 and Kohler 2009a, b for reviews).

More recently, support for rhythmic classes has been said to come from rhythm metrics, such as the Pairwise Variability Indices (Grabe & Low 2002) and the %V-ΔC combination proposed by Ramus, Nespor & Mehler (1999). The aim of these metrics is to measure the variability of vocalic and consonantal intervals in speech and to use the related scores in order to place a language in one of the rhythm classes.

Despite their initial success, metrics have been shown for some time to be problematic as measures of rhythm. Grabe & Low (2002), who examined a sample of 18 languages and computed both PVIs and %V-ΔC for their samples, found that each set of metrics classified some languages differently (e.g. Thai was classified as syllable-timed by %V-ΔC, but as stress-timed by PVIs). More recently, Arvaniti (2009) and Arvaniti & Ross (2009) tested various metrics (including %V-ΔC, and PVIs) by examining (stress-timed) English and German, (syllable-timed) Italian and Spanish and (unclassified) Greek and Korean and eliciting data from eight speakers of each language in three different conditions (isolated sentences, story reading and spontaneous speech). Metric scores showed great inter-speaker variability which increased dramatically in spontaneous speech: as a result, inter-language differences in scores were minimal, especially in spontaneous speech. Further, the scores depended to a large extent on the materials used so that materials incorporating more variability in syllable structure yielded higher metric scores, while materials deliberately designed to avoid such variability yielded lower scores. As a result, the classification of languages shifted from stress- to syllable-timing depending on the materials used to calculate scores. Overall, the results of Arvaniti (2009) and Arvaniti & Ross (2009) show that metric scores can vary in unpredictable ways, suggesting both that metrics are not robust measures and that their variability is opaque and thus difficult to control. These results add to an increasingly large body of research which suggests that metrics cannot rhythmically classify languages or provide insight into the nature of linguistic rhythm even when large samples and a large number of languages are involved (among many, Barry, Andreeva & Koreman 2009, Loukina et al. 2009; see Arvaniti 2009 for a review).

Yet the notion of rhythm classes remains strong and has been used to support language acquisition (among many, Nazzi, Bertoncini & Mehler 1998, Nazzi, Jusczyk & Johnson 2000) and speech processing (e.g. Cutler et al. 1986, Culter & Otake 1994, Mutry, Otake & Cutler 2007). Thus, it is worth considering alternative bases for the notion of rhythm classes that do not rely on timing relations in production. An obvious basis could be found in perception, since "rhythm typology has its roots in auditory observation" as noted by Barry et al. (2009), and in particular on the impressions that different languages, notably English and French (Lloyd James 1940) and English and Spanish (Pike 1945), gave to trained listeners.

Despite the obvious need for exploring perceptual aspects of rhythm, perception studies have been few and far in between and have yielded mixed results. Early studies, such as that of Scott, Isard & de Boysson-Bardies (1985) showed that English and French participants behaved very similarly in a tapping task involving both French and English stimuli, suggesting that listeners' responses to rhythm may not be influenced by either their native language or the language of the stimuli. On the other hand, Miller (1984) found little evidence that phonetically trained and naïve listeners can classify languages into stress- and syllable-timing, but she did find differences depending of the listeners' native language: e.g. French participants (with or without phonetic training) classified Spanish as stress-timed while English participants did not. More recently, Ramus, Dupoux & Mehler (2003) tested listeners' ability to discriminate between English, Dutch, Spanish, Catalan and Polish using impoverished signals that lacked intonational information but kept the timing patterns of the original sentences. They found that some pairs of languages said to belong to the same rhythm class, e.g. English and Dutch, were more difficult to discriminate than pairs across the rhythm class divide, such as English and Spanish, and concluded that their data supported the rhythm class hypothesis.

However, the results of these studies are hardly conclusive and may also be problematic. The task used by Scott et al. (1985) may have shown little differentiation between English and French participants and stimuli because the subjects were not directly asked to perform a rhythm-related task; rather, participants were asked to tap when they heard words that began with [d], with [d]s being evenly spaced in both the French and the English stimuli. Miller's naïve subjects may have found the explicit instruction to classify languages into stress- or syllable-timed too difficult, while Miller herself admits that phoneticians taking part in her experiment may have been influenced by their training. Finally, the results of Ramus et al. (2003) are not all compatible with the idea of rhythm classes. Specifically,

Ramus et al. found that Polish – which Ramus et al. (1999) classified as stress-timed – is discriminated from both stress-timed English and syllable-timed Spanish. Results showing that languages said to belong to the same rhythm class can be discriminated on the basis of degraded signals has been reported elsewhere as well, e.g. for English and Arabic (Mofta & Roach 1988).

The poor results of such experiments suggest that new protocols may be needed to test the idea of distinct rhythm classes. Such protocols should go beyond simple discrimination (which could be due to a variety of confounding factors) and should be neither too indirect, like the tapping task of Scott et al. (1985), or too explicit, like the categorization task of Miller (1984).

In the present study we tried to address precisely these shortcomings by exploring an idea implicit in the rhythm classes, namely that stress-timed languages have a rhythm that is akin to a series of trochees with one prominent syllable followed by less prominent material within each foot, while syllable-timed languages have a rhythm more akin to a simple cadence. If so, then listeners should be able to rate utterances from stress-timed languages as more similar to a series of trochees than utterances from syllable-timed languages. In addition, since Miller (1984) has shown that rhythm classification can be influenced by the native language of the listeners, we also wanted to explore the extent to which our listeners' ability to perform this task and their ratings of different languages would depend on their native tongue. In particular, we expected English listeners to be more attuned to differences in prominence among syllables and more accustomed to regularly occurring prominences, and thus less likely to rate the stimuli as trochee-like. We expected the opposite from Greek participants, who speak a language with strong stresses but tolerate irregularities in prominence patterns much more than speakers of English do (Arvaniti 2007), and we expected Korean participants to find the test more difficult than the other two groups since their native language lacks stress altogether (Jun 2005). Finally, we expected all three groups to find stimuli of their language more rhythmical than those of other languages.

2. Method

To test the above hypotheses, we conducted a perception experiment in which listeners used a scale from 1 to 7 to rate the similarity between low-pass filtered sentences from different languages and a sequence of non-speech trochees.

2.1. Stimuli

The stimuli were sentences recorded for the production experiment of Arvaniti & Ross (2009). Specifically, 12 sentences of each of the six languages in that study (English, German, Greek, Italian, Korean, Spanish) were selected from the data of two native speakers of each language, one male and one female (i.e. there were six sentences from each speaker). The speakers and sentences were chosen by the second author using fluency as the sole criterion for selection. The sentences of each speaker were divided into three pairs: "stress-timed" sentences that is, sentences devised to show as much consonantal and vocalic interval variability as possible, "syllable-timed" sentences, devised to show as little interval variability as possible, and uncontrolled sentences, selected from original works of the languages in question (literary works were used for English, Spanish, Greek and Korean, and text books for German and Italian). The English sentences can be seen in Table 1. The sentences were low-pass filtered at 450 Hz, a level that was tested with several listeners prior to the experiment to ensure that the prosodic features of the sentences were audible but the language could be not inferred from the signal. After filtering, care was taken to adjust the amplitude of the stimuli so that they were comparable in loudness both to each other and to the trochee series.

The non-speech trochees were created using the MacOSX "frog" sound repeated twice with 125 ms of silence between repetitions. In order to make the sequence sound like a trochee, the second repetition was made shorter and less loud than the first: the first repetition was 260 ms long while the second one was 120 ms; the mean intensity of the loudest part of each "frog" was 70 dB for the first repetition and 66 dB for the second. This trochee was heard four times with 220 ms of silence between repetitions, for a sequence total of 2.68 s. Prior to the experiment, the trochee series was tested to ensure that it did not sound either fast or slow, as we wished participants to focus their comparison on the *rhythm* rather than the *tempo* of the trochees and the stimuli.[1]

Table 1: The English sentences used as stimuli

Sentence type	Sentences
"stress-timed"	*The problem required quite a lot of strange equations and wasn't very easy.* *The production increased by three fifths in the last quarter of 2007.*
"syllable-timed"	*Lara saw Bobby when she was on the way to the photocopy room.* *Two-year-old Lucy has macaroni and cheese every day for diner.*
"uncontrolled"	*It was nine o'clock when we finished breakfast and went out on the porch.* *Some little boys had come up on the steps and were looking into the hall.*

The stimuli were collated using PRAAT to create one sound file (an illustration is available at http://idiom.ucsd.edu/~arvaniti/SP2010/practice.wav). Two randomization orders were prepared and counterbalanced across participants within each language group. There was a 1 s silent interval between the trochee series and following sentence, and a 2 s silent interval between trials (see Figure 1). Each trochee-stimulus sequence was heard once for a total of 72 trials. The experiment lasted approximately 10 minutes.

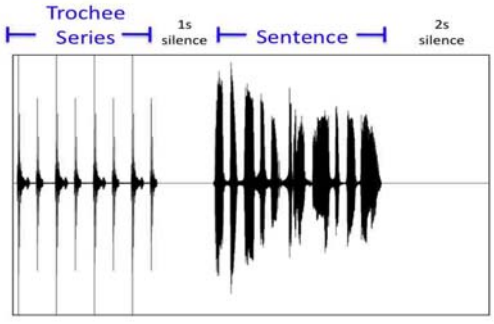

Figure 1: *Timeline of a trial.*

[1] Eight repetitions of the "frog" sound all having the same duration and amplitude were used to create a cadence, but initial results confirmed that, as expected, listeners heard this sequence as a series of trochees (Fraisse 1982), so this option was not pursued further.

2.2. Listeners

Three groups of listeners took part in the study; 22 of them were native speakers of Southern California English, 21 were native speakers of Seoul Korean and 23 were native speakers of Standard Athenian Greek. The first two groups were tested at UCSD and were recruited from among the UCSD population. The average age was 20.75 for the American speakers and 23 for the Korean speakers. All took part in the experiment for course credit, except for five Korean speakers who were graduate students and were paid a small fee for their participation. The Greek speakers were recruited in Athens, Greece and were all volunteers. They were also older than the participants in the other two groups ($\bar{X}$ = 36.5 years old) and the vast majority had at least a Master's degree. All the participants were naïve as to the purposes of the experiment. None of the Korean speakers had any training in linguistics in general or phonetics in particular. Two of the English speakers had some undergraduate training in linguistics, while six of the Greek participants had a Master's or Ph.D. in linguistics, with two of these having a specialization in phonetics.

2.3. Procedures

The English and Korean listeners were tested in a sound-treated room in the UCSD Phonetics Laboratory. The Greek listeners were tested in a quiet room at the Institute for Language and Speech Processing (ISLP) in Athens. All participants heard the experiment through headphones using the facilities of PRAAT and a laptop.

The participants were provided with an answer sheet that included instructions in their native language. They were asked to compare each sentence to the non-speech trochee series in terms of rhythm. Participants were warned that the rhythm of the sentences was unlikely to be identical to the trochees, but they were asked to provide a rating even if uncertain. A scale of 1 (most dissimilar) to 7 (most similar) was used for rating in order to ensure that answers would be spread over the scale as much as possible. The participants were asked to circle the number that was closest to their impression of each stimulus.

Testing was preceded by a short practice session that included five trials and used sentences that were not included in the experiment but were selected from the data of the same speakers.

Many participants did not find the task easy but the vast majority were able to complete it without problems. A small number of questionnaires had to be discarded because participants did not provide a response to some stimuli; this applied to two English and two Greek participants. In addition, one Greek participant did not complete the experiment because she was called away for work. Finally, two Korean participants were excluded because it was discovered after they had completed the test that they had arrived in the US well before adolescence. This brought the number of usable answer sheets to 20 for English, 19 for Korean and 20 for Greek, for a total of 59.

2.4. Statistical analysis

The results were statistically analyzed by means of Analysis of Variance (ANOVA). Preliminary analysis showed that randomization order was not significant ($F < 1$), so this factor was not included in subsequent analyses. The data were analyzed using a mixed design ANOVA with stimulus language as a repeated-measures factor with six levels (English, German, Greek, Italian, Korean, Spanish), sentence-type as a repeated-measures factor with three levels ("stress-timed", "syllable-timed", "uncontrolled"), the participants' language as a categorical predictor with three levels (English, Korean, Greek), and ratings as the dependent variable. An additional ANOVA was run with stimulus language and stimulus speaker as repeated-measures factors and the participants' language as a categorical predictor to see if some of the speakers were judged to be more rhythmical than others.

3. Results

The statistical analysis showed that the listener's native language did not have a significant effect on responses [$F < 1$] and did not interact with stimulus language or sentence-type. The same applied to stimulus speaker, which did not have an effect on responses [$F < 1$] and did not interact with stimulus language or participant's language, a result clearly showing that the ratings were not driven by individual speaker differences in rhythmicity.

Responses, however, did differ depending on the language of the stimuli [$F(5,280) = 2.7$, $p < 0.02$]. Pair-wise planned comparisons showed that differences were due only to English stimuli, which were judged less similar to a trochee than those of any other language [$p < 0.01$ in all cases except English vs. Greek, where $p < 0.04$]. There were no differences between any other languages (see Figure 2), but overall Italian was judged most rhythmical and English least rhythmical in the set.

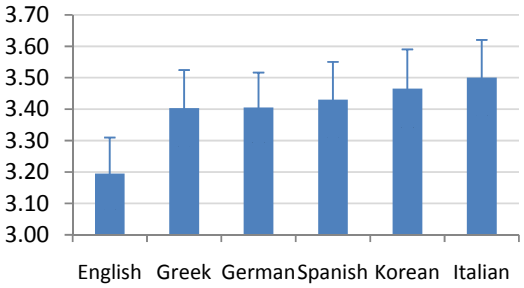

Figure 2: *Mean ratings and standard errors for each language tested.*

Sentence type also affected responses [$F(2,112) = 5.9$, $p < 0.004$]. Pair-wise planned comparisons showed that "uncontrolled" stimuli were rated more trochee-like than "syllable-timed" stimuli [$F(1,56) = 11.98$, $p < 0.001$]. "Uncontrolled" stimuli were also rated more trochee-like than "stress-timed" stimuli, though the difference narrowly missed significance [$F(1,56)=3.9$, $p < 0.054$]. On the other hand, the ratings were not different for "stress-timed" vs. "syllable-timed" stimuli.

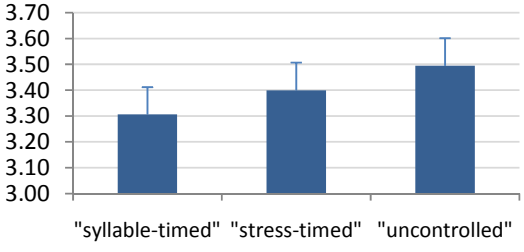

Figure 3: *Mean ratings and standard errors as a function of sentence type.*

4. Discussion and conclusions

Overall, the data show that none of the languages was considered to be very similar to trochees by any group of listeners. It is possible that this was due to the simplicity of the trochee pattern, which did not match the rhythm of any language. Yet it is still surprising that English was judged by all groups of listeners to be the language with the least trochee-like rhythm. This goes against the prediction that English, the quintessential stress-timed language, would be rated most trochee-like. It is not even possible to suggest that the different rating of English reflects its different rhythm class. If that were so, German too would have been rated less trochee-like than the other four languages; but German, as shown, was not rated differently from Greek, Italian, Korean or Spanish.

It is also worth noting that, overall, "uncontrolled" stimuli were rated as more trochee-like than the rest. It is not clear what this result is due to, but it is possible that the "uncontrolled" sentences, being crafted by professional writers, were more natural and therefore they were read more fluently, a difference that could have contributed to their higher rating. On the other hand, the lack of difference in the ratings of "syllable-timed" and "stress-timed" stimuli casts serious doubt on the notion that rhythm class affiliation is directly related to syllable complexity and the auditory impression such complexity gives: if this were so, "stress-timed" stimuli would have received higher ratings than "syllable-timed" ones.

The results were also surprising in that they did not show any differences in the performance of the three groups of listeners as expected on the basis of previous studies, such as Miller (1984). It is possible that the background of the speakers had some effect on the results, especially for the Korean group: all our Korean participants had spent several years in the US and had extensive contact with English. This explanation, however, does not hold for the Greek group: none of the Greek participants had extensive contact with English, and Greek and English are rhythmically distinct (Arvaniti 2007). It is also worth noting that the listeners did not rate the stimuli from their native language any differently from those of other languages. Most strikingly, English-speaking participants rated the English stimuli as the least trochee-like in the experiment. Overall then, the results indicate that all listeners found the task difficult (and possibly too indirect), and thus responded in a similar manner. This idea is also supported by the fact that the participants by and large did not utilize the edges of the rating scale. These two findings together strongly suggest that rhythmic classification on the basis of auditory impression is not easy. If so, the role that rhythm classes play in acquisition becomes questionable.

In conclusion, the present study shows that language classification by means of rhythmic classes cannot be achieved on the basis of listener impressions any more than it can rely on measuring consonantal and vocalic variability in production. Thus, these data add to the small set of perceptual studies of rhythm, and like previous studies they cast doubt on the notion of rhythm classes. The idea of rhythm classes is not supported by either production or perception data so far and this inability to empirically document it renders imperative the search for alternative theories and protocols on which to base the investigation of speech rhythm.

5. Acknowledgements

We would like to thank our participants, especially the ISLP members who generously contributed their time. Special thanks are due to Marianna Katsoyannou who arranged the first author's visit to ISLP, to Jini Shim, Noah Girgis, Christina Lee and Amanda Simons for technical and clerical support, to Julian Bauman for crucial volunteer help, to Jennifer Cole, Chilin Shih, José Hualde, Suzanna Fagyal and their students at UIUC for valuable feedback on this study, and to the many Phonetics lab members who listened to the stimuli and provided input. The financial support of the UCSD-COR through grant LIN210G is hereby gratefully acknowledged.

6. References

Arvaniti, A. (2007). Greek phonetics: The state of the art. *Journal of Greek Linguistics, 8*, 97-208.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66*, 46-63.

Arvaniti, A. & Ross, T. (2009). Rhythm metrics explain little about speech timing and rhythm. UCSD manuscript.

Barry, W., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica, 66*, 78-94.

Bertinetto, P. M. (1989). Reflections on the dichotomy <<stress>> vs. <<syllable timing>>. *Revue de Phonétique Appliquée, 91*, 99-129.

Cutler, A., Mehler, J., Norris, D. & Seguí, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385-400.

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language, 33*, 824-844.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149-180). New York: Academic Press.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515-546). Berlin: Mouton de Gruyter.

Jun, S. (2005). Korean intonational phonology and prosodic transcription. In S. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 201-229). Oxford: Oxford University Press.

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

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

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

Loukina, A., Kochanski, G., Shih,C., Keane, K. & Watson, I. (2009). Rhythm measures with language-independent segmentation. *Proceedings of Interspeech 2009, September 6-10, 2009*. Brighton, UK.

Miller, M. (1984). On the perception of rhythm. *Journal of Phonetics, 12*, 75-83.

Moftah, A. & Roach, P. (1988). Language recognition from distorted speech: comparison of techniques. *Journal of the International Phonetic Association, 18,* 50-52.

Murty, L., Otake, T., & Cutler, A. (2007). Perceptual tests of rhythmic similarity: I. Mora rhythm. *Language and Speech, 50*, 77-99.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology, 24*, 756-766.

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

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

Ramus, F., Dupoux, E. & Mehler, J. (2003). The psychological reality of rhythm class: Perceptual studies. In *Proceedings of the 15th ICPhS*, Barcelona (pp. 337-340).

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265-292.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics, 13*, 155-162.