Testing the perception of speech rhythm on natural and artificial stimuli

Paolo Mairano [1], *Antonio Romano* [2]

[1] GIPSA-Lab, University of Grenoble 3, France
[2] LFSAG, University of Turin, Italy
paolo.mairano@gipsa-lab.grenoble-inp.fr, antonio.romano@unito.it

Abstract

It has long been assumed that the stress-timed vs. syllable-timed dichotomy is based on perceptual impressions of speech rhythm. However, experimental tests are rare in the literature and not all of them have successfully found perceptual evidence for speech rhythm categorization. Experimental protocols have been very different, sometimes testing naïve vs. non-naïve listeners, sometimes using natural speech stimuli, sometimes preferring synthetic stimuli, sometimes using filters to hide lexical information from speech.

In this paper we describe the results of a perceptive test that has been carried out on 43 Italian listeners, who were asked to categorize both natural and artificial stimuli. Results are highly controversial: listeners do not seem to be able to categorize artificial stimuli reproducing prosodic cues of different languages. Yet, there is a mild tendency on the part of speakers to categorize natural speech stimuli of unknown languages in a way that seems to reflect rhythm classes.

Index Terms: speech prosody, speech rhythm, perceptive test, natural stimuli, artificial stimuli.

1. Introduction and theoretical background

The traditional dichotomy of stress-timed vs. syllable-timed languages has been proposed by [1] and [2]. The latter suggests that stresses are roughly isochronous in English, Russian and Arabic (hence stress-timed languages), while syllables tend to be isochronous in French, Telugu and Yoruba (hence syllable-timed languages). This dichotomy (see [3] for further details) is rooted in perception and has enclosed perceptual evaluations even in denominations such as *machine-gun rhythm* (referring to syllable-timed languages) vs. *Morse code rhythm* (referring to stress-timed languages) by [4]. The existence of this impression was usually confirmed even by authors who set out to look for isochrony and who did not find it in acoustic measurements: "a language is syllable-timed if it *sounds* syllable-timed" ([5]:78).

Given such a widespread consensus that syllable-timed languages sound syllable-timed and stress-timed languages sound stress-timed, one would expect that a number of experimental tests have confirmed these claims. However, very few authors set out to verify precisely to what degree languages are perceived as belonging to different rhythm categories.

[6] carried out a test on naïve and non-naïve listeners. Participants were asked to rate non-masked and non-manipulated samples of read speech (*The North Wind and the Sun*) and spontaneous speech as either syllable-timed or stress-timed. Results showed that phoneticians' ratings were unsurprisingly more consistent with expectations than naïve speakers' ratings. Moreover, Arabic was nearly universally perceived as stress-timed, while Indonesian, Yoruba and Japanese tended to be classified as syllable-timed; a higher level of indecision was found for Finnish, Polish and Spanish (but it has to be noticed that the classification of Spanish as syllable-timed has been very controversial, see for instance [7]).

More recently, Ramus, Mehler and co-workers have developed a new experimental protocol for testing the perceptual categorization of speech rhythm that involves de-lexicalizing speech to prevent listeners from rating stimuli on the basis of lexical information. This is achieved through a re-synthesis of the original speech samples in a degraded signal called *flat SASASA*, obtained by resynthesizing all consonants as [s], all vowels as [a] and by leveling pitch and intensity. [8] used the *flat SASASA* synthesis for discrimination tests of English vs. Spanish, English vs. Dutch, Polish vs. English, Polish vs. Spanish, Catalan vs. English, Catalan vs. Spanish and Polish vs. Catalan. Results confirmed expectations: listeners were found to discriminate languages belonging to different rhythm classes (English vs. Spanish), but not languages of the same rhythm class (English vs. Dutch and Spanish vs. Catalan).

SASASA tests were also carried out by [9] to check whether the results would correlate with the values of varcos and %V for the same samples (SSBE having high varcoC and low %V, Orkney Islands and Welsh Valleys English having medium values of both measures, Castilian Spanish having low varcoC and high %V). Data included controlled sentences by 3 English speakers (Welsh Valleys, Orkney Islands and SSBE) and 4 Castilian Spanish speakers. Listeners were found to discriminate Castilian Spanish vs. the three types of English, but not Orkney Islands and Welsh Valleys English.

[10] carried out a test to verify whether speech rate plays a role in the perception of rhythm classes. Participants had to listen to de-lexicalized stimuli of "syllable-timed German and stress-timed French" and to rate them on a scale of regularity. They were unaware that they were listening to manipulated speech samples. Results showed that listeners generally rated stress-timed French samples as being more regular than syllable-timed German samples: this seems to prove that they did not use the variability of vocalic and consonantal intervals as a cue of regularity. Instead, the author suggests that they used the CV rate (the number of vocalic and consonantal intervals per second), which is confirmed by the linear regression in cross-plots of listener ratings of regularity in function of CV-rate.

[11] claimed 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 […] nor too explicit" ([11]:2/4). They built a test in which stimuli were obtained by low-pass filtering sentences of English, German, Greek, Italian, Korean and Spanish at 450 Hz. Sentences of each language were divided into 3 types (syllable-timed, stress-timed, uncontrolled). Listeners (of three different mother tongues, namely English, Greek and Korean) listened to a synthetic trochee series and to a sentence, repeating this task for each sentence. They were asked to rate the similarity of each stimulus to the trochee series on a 7-step

scale. Final results show that the native language of speakers did not significantly affect the ratings, and that stimuli of English were rated less similar to the trochee series. The three sentence types were rated as more similar to the trochee series along the following scale: syllable-timed - stress-timed - uncontrolled. This is at odds with expectations and the authors conclude that "language classification by means of rhythmic classes cannot be achieved on the basis of listener impressions anymore than it can rely on measuring consonantal and vocalic variability in production" ([11]:4/4).

Several researchers have worked on the categorization of languages with rhythm metrics, %V, deltas (see [12]) and PVIs (see [13]) among others. Yet, these acoustic results have rarely been compared with perceptive data, so that whenever there is a discrepancy between the researcher's prediction and the actual results, it is not clear whether this has to be attributed to a malfunctioning of the metrics or to an incorrect auditory impression.

So, we decided to carry out perceptive tests in order to give a contribution to the study of speech rhythm perception by checking both natural and masked samples. We also aimed to compare perceptive data with data on rhythm metrics published in our previous studies (e.g. [14]). After all, [12] intended their measures to be "correlates of the perception of rhythm" and specified that their study was "meant to be an implementation of the phonological account of rhythm perception" ([12]:274). The authors themselves presented the results of a series of tests carried out on adults and infants on the discrimination of languages on the basis of rhythm.

2. The test

We administered a perceptive test to 43 listeners, mainly BA students at the University of Turin. Age range was between 19 and 60 years averaging 25.25. 37 participants had Italian as their mother tongue, 2 had French, 1 had English, 1 had German, 1 had Romanian, 1 had Arabic. No one claimed suffering from hearing impairments.

The test consisted of 4 parts, each lasting 7-10 minutes. Only parts 2 and 4 are reported here as parts 1 (discussed in [15]) and 3 were meant to test other subjects (the perception of lexical stress and of stimuli with manipulated pitch and duration) and are therefore not relevant for this study.

2.1. Testing artificial stimuli

Part 2 of the test was meant to verify the discrimination of rhythm classes on artificial stimuli. Stimuli were obtained by reproducing a stylized version of the prosodic parameters of the first sentence of *The North Wind and the Sun* in 15 languages. The durations and the values of *fo* and intensity for each original vowel were reproduced in a synthetic periodic waveform, while consonantal intervals were substituted with silence.

The stylization procedure which has been used shares the basic assumptions of the best known close-copy stylization defined by [16]. It provides a synthetic approximation of the natural course of the three prosodic parameters (pitch, duration and intensity for each segment), with the basic criterion that the prosody of the final sample should be perceptually indistinguishable from the original. The values stylized for each original vowel were reproduced in a synthetic periodic waveform (series of pulses), while consonantal intervals were substituted with silence (occasionally broken by isolated pulses representing inner bursts in clusters of obstruents). Stimuli were 4-8 s long.

Participants were told they would listen to masked speech stimuli and were asked to decide which language was being spoken. They had to choose between: (1) *Spanish, French or similar* (2) *English, German or similar* (3) *Other* (4) *I don't know* (the first two possible choices were obviously intended to reflect the two traditional rhythm classes). The interface was extremely simple and intuitive, consisting exclusively of the stimulus label and the four buttons, as can be seen in figure 1. Participants only had to press the button corresponding to their answer and they were immediately put forward to the next audio sample. It was not possible to listen to the audio sample more than once nor to go back and correct the answer once it was given. The format is not so different from *SASASA* tests. In this case, participants did not hear three stimuli (A, B and X), but only one (X), and had to classify it on the basis of categories which they presumably already knew. It could be argued that not everybody has ever heard French, German and English (the test was taken in Italy, which guarantees that everybody was at least familiar with Italian). The answer to this is that, first of all, it is not necessary to know all four languages: it is enough to have heard at least (a) Italian or French and (b) English or German. Secondly, most participants were university students at the faculty of foreign languages, who should then be fairly knowledgeable about languages. Finally, nobody complained that they did not have a sufficient knowledge of these languages in order to complete the task.

The hypothesis is that if rhythm classes are rooted in perception, participants should classify stress-timed languages as *English, German or similar*, syllable-timed languages as *Spanish, French or similar* and mixed languages as *other* or *I don't know*. Such a design does have the drawback that evaluations might in come cases be influenced by some a priori on the part of participants about the 4 languages used as reference, but rhythm classes were still expected to emerge in terms of *general trends*.

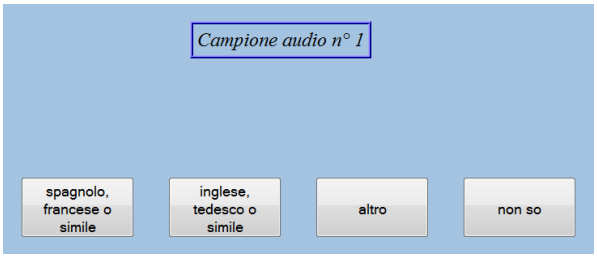

Figure 1: *Screenshot of the interface, consisting merely of a label indicating the stimulus number and the four choice buttons. There is no button to listen to audio samples nor to proceed, as the succession of events is completely controlled.*

2.2. Testing natural speech

The final part of the test consisted in a "scalar" implementation of a traditional ABX categorization test. Participants had to listen to two artificial versions of the first sentence of the *North Wind and the Sun* in RP English (A) and Standard French (B) (but they were not told it was French and English). Then, they had to listen to natural speech samples and to decide whether they resembled more to A or to B. They

had to express their judgment with the help of a slider, which went from A to B (see figure 2).

This procedure was repeated twice: the first time with 7 supposedly unknown languages (Amharic, Czech, Finnish, Standard Belgian Dutch, Icelandic, Indonesian and Turkish), the second time with 7 regional varieties of English (RP English, Tyneside English, New Zealand English, GA English, Australian English, Liverpool English, Southern Michigan English). Samples of Samples of Amharic, Standard Belgian, Indonesian and all English varieties were taken from the *Illustrations of the IPA* (for a complete reference, see http://www.sil.org/~olsonk/ipa.html). In both tasks, the A and B stimuli remained unchanged and participants were free to listen to them as many times as they wished; likewise, they could listen to the 7 samples as many times as they wished and in the order they preferred.

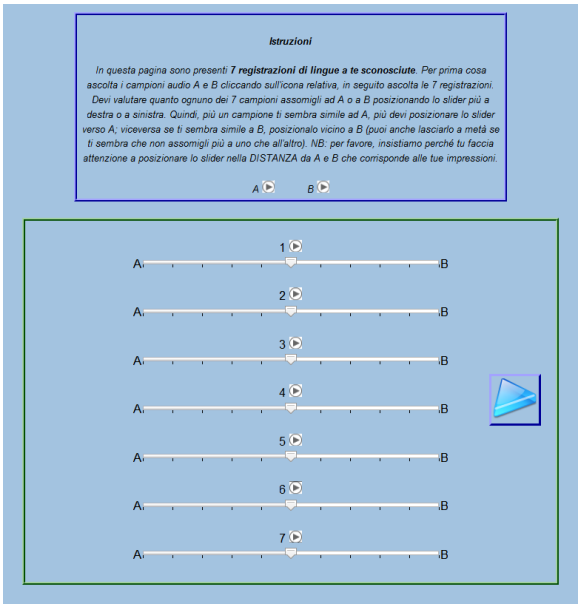

Figure 2: *The interface of the final part of the test. Participants could listen to A, B and the 7 samples by clicking on the corresponding icons. Sliders could be dragged left or right to reflect each sample's resemblance to A and B.*

In short, languages were rated on a continuum: the task did not prompt for a clear-cut bi-polarization, which also allowed them to create an order among the seven unknown languages.

The results

3.1. Artificial stimuli

The results of the categorization of non-speech samples are reported in figure 3. Histograms show the answers given by the 43 participants for each sample. Red bars indicate stress-timed ratings (*English, German or similar*), green bars indicate syllable-timed ratings (*French, Spanish or similar*), blue bars indicate ratings for *Other*, while yellow bars indicate ratings for *I don't know*.

Results are surprising: German, Brazilian and European Portuguese, Romanian, Japanese and French samples display the highest stress-timed ratings, while Russian, Finnish and English (both GA and RP) samples display the highest syllable-timed ratings. High levels of indecision seem to affect the classification of Italian, Icelandic, Turkish, Spanish and Czech.

Even though participants might have been influenced by various linguistic and extra-linguistic factors when judging a language as closer to French/Spanish or to English/German, rhythm classes were expected to emerge as general trends. Instead, it goes without saying that this scenario does not reflect the traditional rhythm classes. It is particularly remarkable that Japanese (supposedly mora-timed), French (supposedly syllable-timed) and German (supposedly stress-timed) are all classified in the same way (namely as *French, Spanish or similar*)!

3.2. Natural speech

The answers given by participants for each of the 14 (7+7) natural speech stimuli are reported in the boxplots in figure 4, showing median values and quartiles. A value of 0 corresponds to stimulus A (stress-timed), whereas a value of 100 corresponds to stimulus B (syllable-timed). Samples pertaining to task 1 (unknown languages) are shown above, while those pertaining to task 2 (varieties of English) are shown below.

The variability of answers is impressing and mostly covers all available space. Median ratings given for unknown languages indicate that Indonesian, Icelandic (and to a lesser extent Flemish and Czech) have been more frequently associated with A (stress-timed), while Turkish, Finnish and Amharic have been more frequently associated with B (syllable-timed).

Results for the second task (regional varieties of English) show very comparable median values for all 7 stimuli, apart from Liverpool English (which seems to have been perceived

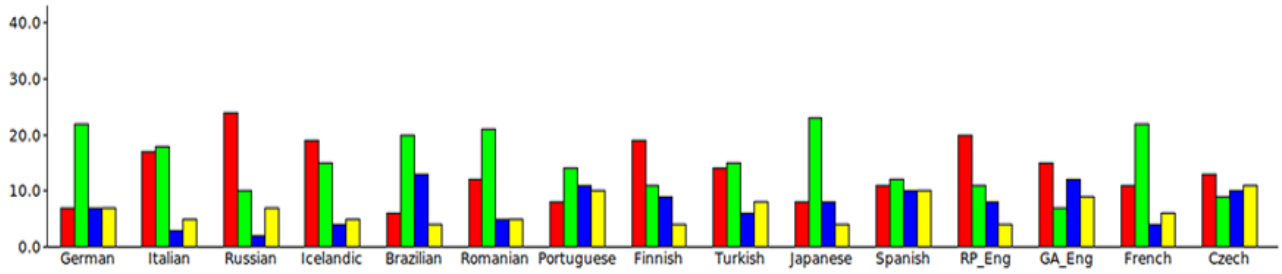

Figure 3: *Answers given by 43 listeners for each masked speech sample. Red bars indicate ratings for "English, German or similar", green bars indicate ratings for "French, Spanish or similar", blue bars indicate ratings for "Other", while yellow bars indicate ratings for "I don't know".*

as slightly more similar to B than the other samples) and RP English (which has been perceived as more similar to A). This is utterly unsurprising, because the artificial A stimulus is precisely a synthesis of that sample of RP English: in other words, listeners merely agreed on the fact that the natural RP English sample sounds like its artificial counterpart. However, despite being uninteresting, this result is reassuring as it indirectly provides a confirmation (1) of the validity of the synthesis method adopted and (2) of the fact that, in general, participants were still concentrating on the test even in its final part and did not simply take wild choices.

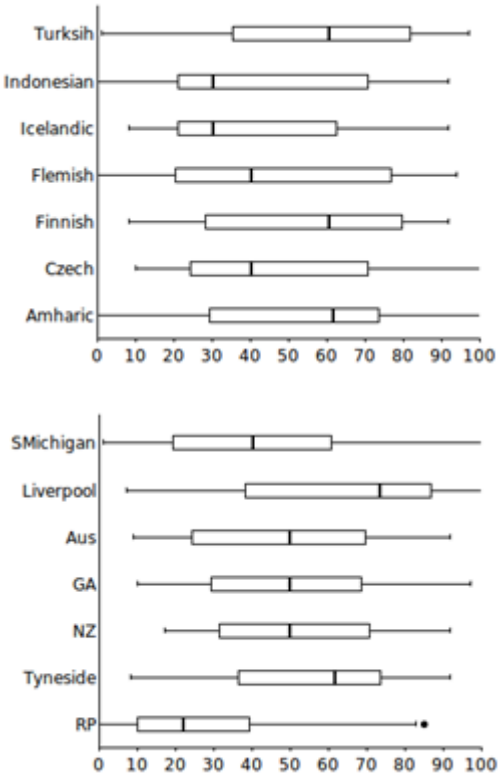

Figure 4: *Results of the final part of the test, asking for a scalar categorization of 7 samples of unknown languages (above) and 7 regional varieties of English (below). Data is presented on box-plots showing the median and quartiles. 0 corresponds to A (artificial RP English sample) and 100 corresponds to B (artificial Standard French sample).*

4. Conclusions

Our data on the perceptual classification of artificial stimuli do not reflect a categorization of speech rhythm classes, with French (supposedly syllable-timed), German (supposedly stress-timed) and Japanese (supposedly mora-timed) rated in the same way (see fig. 3). This therefore confirms the results obtained by [11]. Instead, perceptive data obtained on natural speech samples yield mild support of the rhythm class hypothesis.

If we hypothesize that artificial samples were perceived as non-speech samples, it could be suggested that a mild (acoustically inexplicable) impression of isochrony may be brought about only when listening to real speech. In other words, it could be inferred that the impression of isochrony is a byproduct of speech and might correspond to a kind of mental effort to regularize irregularities. However, there are two problems with this hypothesis. Firstly, although artificial samples were clearly non-verbal, it remains to be seen to what extent they were really perceived as not pertaining to speech. Secondly, [11] have already proven that filtered speech samples (more similar to natural speech) are not categorized in compliance with rhythm classes. In the future, we aim to test this hypothesis on more data and with a protocol addressing more specifically this issue.

Moreover, it has to be remarked that the stimuli of our test preserved the original pitch contour, like in [11] but in contrast to the *flat SASASA* approach (see [8]). Since studies working with flat *SASASA* seem to provide more consistent results, one could hypothesize that pitch *fo* disturbs more than helps listeners in categorization tasks: this aspect needs further clarification.

5. References

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

[2] Abercrombie, D. (1967) *Elements of General Phonetics*, Edinburgh: University Press.

[3] Barbosa, P.A. (2000) Tempo-silábico em Português do Brasil: a critic to Roy Major, *D.E.L.T.A.*, 16/2, 369-402.

[4] Lloyd James, A. (1940) *Speech signal in telephony*, London: Pitman & Sons.

[5] Roach, P. (1982) On the Distinction between 'Stress-timed' and 'Syllable-timed' Languages. In D. Crystal (ed.), *Linguistic controversies*, London: Edward Arnold, 73-79.

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

[7] Borzone de Manrique, A. M. & Signorini, A. (1983) Segmental duration and rhythm in Spanish, *Journal of Phonetics*, 11, 117-112.

[8] Ramus, F. Dupoux, E. & Mehler, J. (2003) The psychological reality of rhythm class: perceptual studies. In: Proc. of the 15th ICPhS, Barcelona (Spain), 3-9 August 2003, 337-342.

[9] White, L., Mattys, S.L., Series, L., & Gage, S. (2007) Rhythm metrics predict rhythmic discrimination. In: Proc. of the 16th ICPhS, Saarbrücken (Germany), 6-10 August 2007, 1009-1012.

[10] Dellwo, V. (2008) The role of speech rate in perceiving speech rhythm. In: Proc. of Speech Prosody 2008, Campinas (Brazil), 6-9 May 2008, 375-378.

[11] Arvaniti, A. & Ross, T. (2010) Rhythm classes and speech perception. In: Proc. of Speech Prosody 2010, Chicago (USA), 11-14 May 2010.

[12] Ramus, F., Nespor, M. & Mehler, J. (1999) Correlates of Linguistic Rhythm in the Speech Signal, *Cognition*, 73/3, 265-292.

[13] Grabe, E., & Low, E. L. (2002) Durational Variability in Speech and the Rhythm Class Hypothesis. In: C. Gussenhover & N. Warner (eds), *Papers in Laboratory Phonology*, 7, Berlin: Mouton de Gruyter, 515-546.

[14] Romano, A. & Mairano, P. (2010) Speech rhythm measuring and modelling: pointing out multi-layer and multi-parameter assessments. In: M. Russo (ed.), *Prosodic Universals: comparative studies in rhythmic modeling and rhythm typology*, Rome: Aracne, 79-116.

[15] Romano A. & Mairano, P. (2011) Prominenze accentuali di frasi italiane nella percezione di un gruppo di studenti torinesi. *Proc. of the 7th AISV Congress*, Lecce, 26-28 Jan 2011.

[16] 't Hart, J., Collier, R. & Cohen, A. (1990) *A perceptual study of intonation*, Cambridge: University Press.