Loquens 1(1)
January 2014, e008
eISSN 2386-2637
doi: http://dx.doi.org/10.3989/loquens.2014.008

Listeners may rely on intonation to distinguish languages of different rhythm classes

Lea Hagmann[1] and Volker Dellwo[2]

[1]Center for the Study of Language and Society (CSLS), University of Bern
[2]Department of Comparative Linguistics, University of Zurich
e-mail: lea.hagmann@iash.unibe.ch, volker.dellwo@uzh.ch

Citation / Cómo citar este artículo: Hagmann, L., & Dellwo, V. (2014). Listeners may rely on intonation to distinguish languages of different rhythm classes. *Loquens, 1*(1), e008. doi: http://dx.doi.org/10.3989/loquens.2014.008

ABSTRACT: Previous research argued that listeners can distinguish between languages of different rhythm class but not of the same class (class discrimination hypothesis). In the present research we tested the role of duration and pitch cues (intonation) in this process. In Experiment I we tested whether we could replicate previous findings on listeners' language discrimination ability with native Swiss German listeners. Results showed that the discrimination of English and Japanese based on durational cues led to the same results as in previous experiments. In Experiment II we tested listeners' ability to distinguish between languages belonging to different rhythm classes (English-French, French-Japanese, Spanish-Japanese) and the same rhythm class (Spanish-French). Results revealed that listeners' distinction was not above chance level for all language contrasts. In Experiment III we added intonation to a French-English and a Spanish-French language contrast. Results revealed a significant effect of intonation for the French-English but not the Spanish-French contrast. The experiments showed that the primary cue for listeners to distinguish between languages of different rhythm class is not generally duration, as previously hypothesized, but it can also be intonation. Implications of the findings on the theory that languages can be classified according to their speech rhythm (rhythm class hypothesis) are discussed.

KEYWORDS: rhythm; intonation; typology; perception

RESUMEN: *Los oyentes podrían basarse en la entonación para distinguir lenguas de diferentes clases rítmicas*.- Algunas investigaciones anteriores sostienen que los oyentes pueden distinguir entre lenguas de diferente ritmo pero, en cambio, no de la misma clase rítmica (hipótesis de la discriminación de clases). En la presente investigación examinamos el papel de la duración y de las claves tonales (entonación) en este proceso. En el Experimento I analizamos si podíamos replicar los resultados anteriores sobre la capacidad de discriminación lingüística de los oyentes con jueces nativos de alemán de Suiza. Los resultados muestran que la discriminación de inglés y japonés basada en claves de duración conduce a los mismos resultados que en experimentos anteriores. En el Experimento II analizamos la capacidad de los oyentes para distinguir entre lenguas pertenecientes a diferentes clases rítmicas (inglés-francés, francés-japonés, español-japonés) y a la misma clase rítmica (español-francés). Los resultados pusieron de manifiesto que la distinción por parte de los oyentes no se encontraba por encima del nivel del azar para todos los contrastes entre lenguas. En el Experimento III añadimos la entonación a los contrastes entre francés e inglés y entre español y francés. Los resultados revelan un efecto significativo de la entonación para el contraste francés-inglés pero no para el contraste español-francés. Los experimentos muestran que la clave primaria que los hablantes usan para distinguir entre lenguas de diferente clase rítmica no es generalmente la duración, como previamente se había propuesto, sino que también puede ser la entonación. Por último, se analizan las implicaciones de los resultados para la teoría de que las lenguas pueden clasificarse según su ritmo de habla (hipótesis de la clase rítmica).

PALABRAS CLAVE: ritmo; entonación; tipología; percepción

1. INTRODUCTION

By now there is a wide body of evidence revealing that listeners have a considerable ability to distinguish between languages based on supra-segmental cues (Ramus, Nespor, & Mehler, 1999; White, Mattys, & Wiget, 2012). A widely held view is that languages can be grouped by their auditory rhythmic characteristics into categories such as stress-, syllable- and mora-timed languages and that listeners can distinguish between languages of different rhythmic classes but not between languages of the same rhythm class (henceforth: class discrimination hypothesis[1]; Nazzi, Bertoncini, & Mehler, 1998; Ramus, Dupoux, & Mehler, 2003; Ramus et al., 1999). While evidence against this view has recently been provided by White et al. (2012) the aim of the present paper was on the one hand to analyze listeners' language discrimination ability for more language pairs to obtain more evidence either for or against the class discrimination hypothesis and on the other hand to find what possible roles duration and intonation might play in this process.

The categorization of languages into rhythm classes harks back to early claims on auditory rhythmic differences between various languages by Arthur Lloyd James (1929) who argued that some languages sounded more regularly timed than others (he coined the metaphors machine-gun and Morse-code rhythm). This resulted in the well known dichotomy of "stress-timed" versus "syllable-timed" languages by Pike (1945) to which a third class "mora-timed" languages was later added (Ladefoged, 1975). Languages have since often been classified as belonging to one of these three different rhythm classes, which possibly turned into one of the most highly disputed hypotheses in phonetics since the 1920s. Connected to the rhythm class hypothesis was the 'isochrony hypothesis' (Abercrombie, 1967), which clearly stated the assumptions towards syllabic and foot durations in syllable- and stress-timed languages respectively. French, for example, has since often been referred to as a canonical syllable-timed language and English as a stress-timed language (Roach, 1982). Since one of the auditory impressions was that syllables in syllable-timed languages reveal a high rhythmic regularity on the syllabic level that is not present in stress-timed languages, it was assumed that the syllable durations must be more equally (quasi-isochronously) distributed in these languages. In stress-timed languages the percept of regularity between stressed syllables was argued to be perceived more regularly, which is why a quasi-isochronous distribution of durations between stressed syllables (i.e., foot intervals) was predicted. Even though the assumptions are possibly not in line with numerous psychoacoustic results (the perception of duration is not independent of factors like sound quality or its intonation) they have more or less been taken for granted in many studies on speech rhythm. However, numerous experiments analyzing syllable and foot durations in a variety of languages revealed that there is no acoustic evidence for the isochrony hypothesis (Dauer, 1983; Ladefoged, 1967; O'Connor, 1965; Ohala, Riordan, & Kawasaki, 1979; Roach, 1982; Shen & Peterson, 1962).

An innovative approach to the problem was proposed towards the end of the 1990s. Based on arguments by Roach (1982) and Dauer (1983) that the perception of regularity and irregularity in syllable- and stress-timed languages might be related to structural differences in phonological/phonotactic characteristics of these languages, Ramus et al. (1999) argued to observe the durational characteristics of consonantal and vocalic interval durations in speech. The rationale for this was that consonantal intervals are typically more complex in stress- than in syllable-timed languages, hence their durational variability should be higher (e.g., as measured by their standard deviation, ΔC; Ramus et al., 1999). Since languages which are argued to be syllable-timed typically do not reveal the phonological feature of vocalic reductions, speakers should spend more of the proportional duration of time during a speech signal on vowels, so their percentage over which speech is vocalic ($\%V$) should be higher. Evidence for these assumptions was provided in Ramus et al. (1999). It led to a large number of studies during the first decade of the 2000s that aimed at further developing the measurement procedures ΔC and $\%V$ (amongst others: Barry, Andreeva, Russo, Dimitrova, & Kostadinova, 2003; Dellwo, 2006). Further, a variety of other languages and dialects were studied (amongst others: Arvaniti, 2012; Dancovičová & Dellwo, 2007; Leemann, Dellwo, Kolly, & Schmid, 2012; Mairano, 2011), which sometimes, in particular in the case of Arvaniti (2012), lead to conflicting results not supporting the rhythm class hypothesis. Alternative approaches in a very similar vein were also present. Grabe and Low (2002), for example, measured the average differences between consecutive consonantal and vocalic intervals (the so called Pairwise Variability Index; PVI).

One of the probably strongest arguments in favor of the rhythm class hypothesis, however, was not the acoustic but the perceptual evidence provided by Ramus and Mehler (1999), Ramus et al. (1999), and Ramus et al. (2003). These studies showed that human adult listeners could distinguish between languages of different rhythm classes (for example English and Japanese, i.e., stress- and mora-timed) but not between languages from the same class (for example English and Dutch, both stress-timed). How was this tested? Ramus and Mehler (1999) followed the rationale they elaborated in Ramus et al. (1999), namely that the durations of consonantal and vocalic intervals contain the key information to a language's rhythmic categorization. Next to the phonological/phonotactic argument (above) the rationale to this view also derived from evidence provided by Nazzi et al. (1998) revealing that listeners (in this case human newborns) are

[1] The terminology was adopted from White et al. (2012).

able to derive rhythmic information from highly low-pass filtered speech (below 400 Hz). This manipulation leaves strong cues to the durations of vocalic intervals in the signal, which, in return, is evidence for the assumption that listeners make use of exactly these cues in language discrimination. Low-pass filtering, however, still leaves rudimentary segmental cues in the signal and in the case of high amplitude voiced consonants (e.g., approximants like /l/ or /w/) the distinction between the vocalic and consonantal parts of the signal is thus not quite clear. In the case of low-pass filtering as applied by Nazzi et al. (1998) also the entire pitch contour was present in the resulting signal. For this reason Ramus and Mehler (1999) developed the so-called "*flat sasasa*," an MBROLA based resynthesis method in which consonantal intervals in speech were turned into /s/ sounds and vocalic intervals into /a/ while a flat (i.e., monotonous) pitch contour was applied to the signal resynthesized sounds. This resulted in a speech-like signal, which contained only cues to the durations of consonantal and vocalic intervals in the speech signal.

Using *flat sasasa*, Ramus and Mehler (1999) tested whether French listeners could distinguish between acoustically modified versions of sentences derived from two exotic languages, "Sahatu" and "Moltec," which were invented names for the real underlying languages English and Japanese. To prevent listeners from using possible acquired knowledge about theses languages in their decision process, the real names to the languages were kept anonymous. Next to durational cues, listeners' performance for a variety of other cues like the phonotactic arrangement of speech and intonation was tested. For this experiment, three more modifications were created: (a) "*saltanaj*": cues to consonant types and vowels as well as the intonation contour were maintained (replacements: all vowels with /a/, all fricatives with /s/, all stops with /t/, all liquids with /l/, all nasals with /n/, and all glides with /j/); (b) "*sasasa*": identical to *flat sasasa* (above) but the intonation contour from the original speech signal was maintained; (c) "*aaaa*": sonorants were replaced with /a/ and linear interpolations between the inter-sonorant-intervals were applied. Given that listeners showed the same (in *saltanaj* and *sasasa*) or worse performance (in *aaaa*) compared to the *flat sasasa* condition, Ramus and Mehler concluded that intonation free durational cues to vocalic intervals are sufficient for French listeners to distinguish between the languages Japanese and English. To us, however, it seems questionable whether this argumentation is justified because in the *aaaa* condition the intonation contour was not only extracted, but additional new and probably distracting intonation information was created at the points where the contour was interpolated. Such interpolation intervals should be longer and more variable in the case of English, where typically a higher number of voiceless consonants can be present between two vowels compared to Japanese. It thus seems plausible that, because of various psychoacoustic effects, numerous cues to intonation were destroyed and possible artifacts were created by this method. This might have lead to the poorer listener performance in the *aaaa* condition. Furthermore, it appears implausible that listeners' performance in language discrimination ability might drop when intonation is provided. Should this be the case, it would mean that listeners cannot rely on the durational information for language discrimination in real speech, as the intonation is always present and would cover the duration cues. This question was addressed in the present paper (Experiment III) when we added intonation to between and within rhythm class language contrasts (French-English, French-Spanish) to test whether this affects listeners' identification performance (see introduction to Experiment III for details). In order to collect more evidence about the class discrimination hypothesis, Ramus et al. (2003) replicated the method in Ramus and Mehler (1999) for a diverse range of within- and between rhythm class contrasts. Table 1 shows the results from these experiments.

Table 1: Mean *A′* values for listener discrimination performance for a variety of language contrasts from Ramus et al. (2003, p. 340).

	A′	St. Dev.	*p*
Exp. 1: English-Spanish	0.65	0.14	0.007[α]
Exp. 2: English-Dutch	0.49	0.11	0.71
Exp. 3: Polish-English	0.59	0.15	0.009
Exp. 4: Polish-Spanish	0.74	0.08	< 0.001
Exp. 5: Catalan-English	0.58	0.13	0.004
Exp. 6: Catalan-Spanish	0.48	0.14	0.42
Exp. 7: Polish-Catalan	0.57	0.15	0.03

It is visible that contrasts between rhythm class (English-Spanish and possibly Catalan-English) are significantly different from each other, and contrasts within category (English-Dutch, Catalan-Spanish) are around the *A′* chance level (0.5) and are nonsignificant. There are a number of other contrasts between languages that are difficult to categorize on an auditory basis (Polish, Catalan, and possibly Spanish; see below), which then again revealed a significant listener discrimination ability when contrasted. In the present study we tested listeners' performance for more language pairs that have traditionally been classified as either stress- or syllable-timed (Experiment II) to find out whether or not there might be further evidence for the class distinction hypothesis. By replicating results from Ramus and Mehler (1999) in Experiment I, we tested whether the outcome of our method was comparable to theirs.

In summary, we carried out the following experiments:

- In **Experiment I** we evaluated whether we can replicate the experiments on the Japanese-English distinction from Ramus and Mehler (1999) with Swiss German instead of French listeners and small modifications in the methodology. This was done to ob-

tain a baseline to make our experimental method directly comparable to previous studies (Ramus et al., 2003; Ramus & Mehler, 1999).
- In **Experiment II** we used our method to test whether Swiss German listeners can discriminate between languages from canonical stress-syllable-timing contrasts, i.e., French and English, based on *flat sasasa*. We further tested other between rhythm class contrasts (French-Japanese, Spanish-Japanese) and a within-class contrast (French-Spanish).
- In **Experiment III** we added intonation to test whether listeners might use these cues to improve their language discrimination accuracy. A between-class contrast (English-French) as well as a within-class contrast (French-Spanish) were tested.

The rationale to the individual studies is further elaborated on in the introductions to the subparts.

2. EXPERIMENT I: THE ABILITY OF SWISS GERMAN LISTENERS TO DISTINGUISH BETWEEN JAPANESE AND ENGLISH IN THE *FLAT SASASA* CONDITION

The first aim of Experiment I was to replicate the results for *flat sasasa* condition in Ramus and Mehler (1999), henceforth R&M, for our particular apparatus, the stimuli and listener groups. Further we applied a small change in the procedure. The major changes in our method that might influence the results are the following:

(a) *Change in listener group*: We tested the experiment with native listeners of Swiss German as opposed to French listeners in R&M. R&M chose French listeners in their design which were asked to judge *sasasa*-speech derived from Japanese and English with the argument that French is argued to reveal rhythmic characteristics (syllable-timing) that are neither like those of Japanese (mora-timing) nor those of English (stress-timing). Our Swiss German listeners speak a native language that would commonly be referred to as stress-timed, i.e., it should be more similar in rhythm to English, rather than to Japanese. It is unclear what the effect of this could be. Given that the listeners might be more familiar with the sound of the rhythm in a language that is rhythmically similar (White & Mattys, 2007), it seems plausible that the English *flat-sasasa*-stimuli sound more familiar to them, and that they should thus be in the position to distinguish between the two types of stimuli even better.

(b) *Change in test procedure*: During the test session in R&M listeners received feedback after responding whether a stimulus derived from a certain language. This may have facilitated the listeners to acquire some knowledge about the test stimuli during the test phase, which might have boosted their performance. We were thus interested in knowing whether the results could be replicated without the presentation of feedback.

(c) *Change in* sasasa-*generation*: In R&M the signals for the *sasasa*-condition were generated using MBROLA speech synthesis. In the present study, we developed a method in which /s/ and /a/ segments are concatenated by an overlap-add method (see method below for details). In order to test this we took the identical sentences from a database by Nazzi et al. (1998) that were used in R&M for Japanese and English, and resynthesized them with our *sasasa*-generator. As the resulting *sasasa*-stimuli sounded very similar to the MBROLA based ones from R&M we did not expect this point to affect our results.

2.1. Method

2.1.1. Subjects

6 native Swiss-German speakers (age range 20 to 30 years) participated in the experiment. All subjects had knowledge of English and French as a foreign language. Subjects were all students at Zurich University and received a small reimbursement in return for their participation.

2.1.2. Material

For each language, Japanese and English, we used 20 different sentences produced by 4 female speakers (5 sentences each) from a corpus recorded by Nazzi et al. (1998). These were the exact same sentence recordings that were used in R&M. Sentences for each language were split into a training and a test set each. For each set the sentences of two randomly chosen speakers per language were chosen. *Sasasa*-delexicalization was carried out with the plug-in "*sasasa*-delexicalizer" written for Praat (Boersma & Weenik, 2014) by the second author[2]. This method resynthesizes recorded samples of /a/ and /s/ produced by a male speaker of German and concatenates them to strings of *sasasa* using an overlap-add method. The sentences created were given a monotone intonation contour at 130 Hertz (the average intonation of the male speaker used for the sound recordings).

2.1.3. Procedure

Before the start of the experiment, sentences in the training and test sets were randomized for each listener. Following the procedure of R&M listeners were told that they would be listening to two distorted exotic languages, Sahatu and Moltec, which they would need to learn to distinguish from each other. In a 2 alternative forced choice design, listeners went through a maximum of three

[2] The plug-in can be found at http://www.pholab.uzh.ch/leute/dellwo/software.html.

training sessions during which they were presented a *sasasa*-version of each sentence in a training set and were asked to respond whether the sentence was Sahatu or Moltec. The stimuli were presented to listeners via headphones on a laptop computer using Praat software. After each presentation of a stimulus, participants saw a screen with two buttons in its center, one labeled "Sahatu," the other "Moltec." Listeners clicked one of the buttons and received written feedback on the screen in large letters telling them whether their choice was correct or not. After the presentation of the feedback, the next stimulus appeared upon a click with the mouse. If listeners performed higher than 75% during one of the training sessions, the training was stopped and listeners went straight into the test phase. At the end of each test session listeners were informed about how high they scored overall.

After the training sessions, listeners went straight into the test session during which the stimuli from the test set were played. They were randomized for each listener. During the test session, listeners were presented the same screen as during the training; however, no feedback followed the choice of language. Instead, the next test stimulus was presented upon a click with the mouse anywhere on the computer screen. The whole experiment took between 10 and 15 minutes per person. After having completed the whole set, participants were asked how they experienced the task and what strategies they had developed in order to tell the languages apart.

2.1.4. Data processing

Listeners' performance was measured as the percentage correct (%*C*) which was the mean of correct identifications for Japanese and correct identifications of English stimuli. As in R&M, we also calculated *A'* from signal detection theory. For this we randomly attributed "target" to English and "noise" to Japanese.

2.2. Results

Listeners' mean %*C* was 74.2, which was clearly above a chance level of 50%. Mean *A'* resulted in 0.81, which, again, was drastically higher than the *A'* chance level of 0.5. The data shows that listeners' sensitivity calculated by *A'* is higher in comparison to %*C*, which means that the poorer performance indicated by %*C* should be due to a listeners' decision bias (i.e., listeners being either biased towards a Sahatu or a Moltec response when unsure). The distribution of the *A'* results confirms the high performance of listeners; the total range was between 0.6 and 0.98. The distribution of the *A'* values can be seen in the box-plot to Experiment II (Figure 1, left plot).

In comparison to the results in R&M where the comparable test group achieved a mean %*C* of 68.1 in the *flat sasasa* condition (R&M, Table 1, p. 515), our test group scored about 6 percentage points higher. The total *A'* range in the R&M group was between 0.25 and 1, which is a drastically larger range than we obtained. This range also shows that a considerable number of subjects scored around chance level (0.5) and lower, meaning that they systematically confounded the languages in the test condition. It is possible that the larger range occurred as a result of the larger group size in R&M ($N = 16$ as opposed to $N = 6$ in our experiment).

2.3. Discussion

We showed that the Swiss German listeners performed equally well if not better for the *flat sasasa* condition involving English and Japanese than the French listeners under the same condition in R&M. Essentially this means that we have successfully replicated the results for the *flat sasasa* condition for Japanese and English in R&M with our method and listener group. It further means that neither our way to process *sasasa* (see point (a) under Section 2 above), nor the change in listeners' native language or the lack of feedback during the test session affected listeners' performance in a negative way. Given that our Swiss German listeners showed some tendency towards a higher performance, our slight modifications rather had a positive effect. It is possible that this is due to the familiarity of Swiss listeners with stress-timing patterns, as hypothesized under (a) above. However, it remains unclear what the actual role of the methodological modifications indicated in points (a), (b) or (c) above was in this possible performance boost. The possibility remains that these changes were interacting. Eventually such slight nuances are irrelevant to the main outcome of the experiment, which is that Swiss German listener groups and our experimental method proved suitable for further testing as they performed similarly to the respective French listener group in R&M.

To what degree did listeners make use of characteristics like %*V* to solve the task? This question is difficult to answer but the interviews we took with subjects after the test gave us some clues for hypotheses. Some participants told us that Sahatu (English) contained a larger number of /s/-sounds, while Moltec (Japanese) had more /a/-sounds. Given that the number of /s/ and /a/ intervals did not vary across the two groups, it is possible that listeners might have been referring to characteristics like %*V* (the proportional time over which speech is vocalic). Some also felt that Moltec sounded much softer than Sahatu. A few listeners mentioned that the rhythm of Moltec seemed more regular than that of Sahatu. These responses might point at the fact that people rely on different criteria to make their decisions which might also explain the variability in their performance from close to chance (*A'* = 0.6) to near perfection (*A'* = 0.98). It seems possible that listeners who mentioned regularity versus irregularity of speech rhythm as their discrimination criteria might have relied on %*V*, as Ramus et al. (1999) suggested in a simulation experiment (p. 279). However, according to this simulation, people should be able to score an average of correct answers of 92.5% in telling Japanese and English apart on the basis of speech rhythm – a number that was neither reached in R&M's experiment nor in the present replication. This would then suggest that only a small number of people relied on %*V* when discriminating between languages. Those who spoke of stronger or more /s/-sounds in Sahatu might have a greater sensitivity for phonetic differences. Their impression presumably derived from the fact that consonant clusters in English appear as long /s/-sounds in resynthesized speech. Given these listener' impressions, it is likely that speech rhythm, and more specifically, duration, did not play the only role in a language discrimination task like the one presented here. Those who spoke of a softer Moltec versus a stronger Sahatu also seem to have based their criteria more strongly on a general impression of the speech samples that might similarly lie in phonetic differences, since the /s/-sound is a much harsher sound than /a/. It is important to see that all the previous interpretations on the use of cues are highly hypothetical mainly for the reason that only a small group of listeners has been tested. We found that this group size was sufficient for replication reasons but can only give ideas about listeners' task-solving techniques.

3. EXPERIMENT II: TESTING THE CLASS DISTINCTION HYPOTHESIS FOR A VARIETY OF LANGUAGES USING *FLAT SASASA*

The class distinction hypothesis (the hypothesis that listeners can distinguish languages between but not within rhythm classes based on durational cues) was challenged in a study by White et al. (2012) where English listeners' ability to distinguish between English and Spanish was compared to their performance in distinguishing between varieties of English (Orkney and Welsh Valleys English). Results based on a slightly modified method (ABX instead of AX) revealed that listeners showed some capacity to distinguish between varieties of the same language, even if this capacity was rather weak. These results were thus the first evidence that listeners can distinguish between languages of the same rhythm class based on durational cues, even between varieties of the same language. The findings gave rise for our Experiment II in which we contrasted language pairs previously not analyzed to investigate listeners' discrimination performance. Both White et al. (2012) as well as Ramus et al. (2003) have contrasted Spanish and English as a between-rhythm class example. The attribution of a language to a certain rhythm class based on auditory characteristics is not without dispute. Given the experimental results and discussions provided in Pointon (1980), Toledo (1988) and Almeida (1997) it is questioned whether Spanish really is a pure syllable-timed language, as it contains both audible and acoustic characteristics of stress- and syllable-timing. This means that with Spanish and English, listeners in both Ramus et al. (2003) and White et al. (2012) have actually revealed that within-rhythm class distinction is possible. For this reason we tested a contrast that has since the earliest times of the rhythm class hypothesis been described as one of the most prominent contrasts, namely English as a stress-timed language in contrast to French as a syllable-timed language (see Lloyd James, 1929). We tested this contrast in Experiment II. Given the choice of these two canonical representatives of their class, we expected listeners' performance to be significantly higher than for previous results of English-Spanish contrasts. On the contrary, when French, a canonical syllable-timed language is contrasted with Spanish for which classification has been argued upon, we would expect equally strong discrimination ability for French-Spanish than we would for English-Spanish. In addition to these two language contrasts, we tested two other between-class contrasts: (a) French and Japanese, another between-class contrast (syllable- and mora-timed) for which we expected a high discrimination ability, and (b) Spanish and Japanese. Given that Spanish might reveal characteristics of a stress- and a syllable-timed language, we assumed that listeners' discrimination ability for this pair should be close to the English-Japanese performance.

3.1. Method

3.1.1. Subjects

48 native Swiss-German speakers (age range 20 to 30 years) participated in the experiment. Subjects were all students at Zurich University and received a small reimbursement in return for their participation. In a between-subject design listeners were randomly attributed to the following test groups: English-French (*N* = 12), French-Japanese (*N* = 12), French-Spanish (*N* = 12), and Japanese-Spanish (*N* = 12).

3.1.2. Material

The stimuli sets were created for each language pair in a way analogous to the stimuli sets described in Experiment I. Sentences for each language were taken from a database used in Ramus et al. (1999).

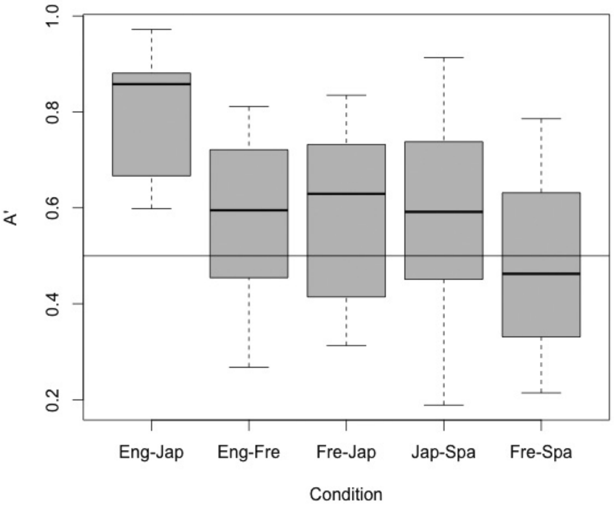
3.1.3. Procedure

The procedure was identical to the test procedure in Experiment I.

3.2. Results

Figure 1 contains the distributions of *A'* for each language-pair group. The figure also contains the results for the listeners group of the English-Japanese condition from Experiment I for comparison reasons. Chance level is at *A'*=0.5. It can be seen that the results for the English-Japanese listening condition from Experiment I were visibly higher than for the other groups.

Figure 1: Box-plots showing the distributions of *A'* (averaged across subjects) for each language-pair condition, Experiments I and II.



To test the discrimination performance of listeners, we carried out five one-tailed *t*-tests testing whether the *A'* group mean is significantly greater than 0.5 (*A'* chance level); one test was performed for each group (Bernoulli corrected alpha level was 0.01; 0.05/5). The results are summarized in Table 2.

Table 2: Results from five one-sample *t*-tests (one in each row) comparing each group's *A'* mean against 0.5 (*A'* chance level).

Group	mean *A'*	df	t	*p*
English Japanese	0.8	5	5.23	0.001**
English French	0.59	11	1.82	0.048
French Japanese	0.59	11	1.59	0.07
Japanese Spanish	0.59	11	1.49	0.081
French Spanish	0.48	11	-0.43	0.66

It is apparent from Table 2 that only the English-Japanese listener group revealed a highly significant performance above chance level. All other listener groups performed insignificantly above chance. The between-rhythm class comparisons (English-French, French-Japanese, Japanese-Spanish) did not reach a significant performance, however their p-level showed some tendency to reach a non-Bernoulli corrected alpha level of 0.05 (in the case of English-French even slightly below). It is possible but unclear whether larger group sizes might have led to significant effects. But also in such cases the magnitude of the effects would have been likely to be very small which means that the performance would need to be interpreted as poor, significant or not. The within-rhythm class comparison, French-Spanish, must be regarded as impossible to solve for the listener group.

3.3. Discussion

The results from this experiment are clear. Swiss listeners that were shown to reveal identical and better performance than their French peers (Ramus & Mehler, 1999) in English-Japanese discrimination based on *flat sasasa* speech failed to do so for typical between-rhythm class contrast English-French and are equally poor on the other between-class contrasts, French-Japanese and Spanish-Japanese. This data shows that Swiss listeners were able to distinguish between some within-class contrasts (English-Japanese) but not between other canonical contrasts like English and French for which our expectation was high. This data thus calls into question the belief that languages reveal duration characteristics of their rhythm class, which listeners might use to distinguish between them. We have provided evidence from three between-class contrasts (English-French, English-Japanese and Japanese-Spanish) for which listeners showed no discrimination sensitivity in *flat sasasa*. This was particularly surprising for the Japanese-Spanish contrast for which we would have expected a high discrimination performance because of the mixed cues to stress- and syllable-timing in Spanish. In summary, our data provided more evidence in support of White et al.'s (2012) view that languages from different rhythm classes cannot be distinguished any better based on durational cues than languages of the same rhythm class; hence our evidence does not support the class distinction hypothesis.

4. EXPERIMENT III: THE INFLUENCE OF INTONATION ON LISTENERS' LANGUAGE DISTINCTION ABILITY BETWEEN AND WITHIN RHYTHM CLASSES

Next to language discrimination tests on human adults (Ramus et al., 2003; Ramus & Mehler, 1999; White et al., 2012), such tests have further been carried out with newborns (Ramus et al., 1999), Tamarin monkeys (Ramus, Hauser, Miller, Morris, & Mehler, 2000; Tincoff et al., 2005) as well as rats (Toro, Trobalon, & Sebastián-Gallés, 2003) and led to the widely accepted conclusion that humans as well as mammals generally seem to be capable of distinguishing languages between but not within rhythm classes. This was taken as evidence for the view that an awareness of the rhythmic organization of speech is present in a pre-linguistic stadium of language acquisition (or non-linguistic in animals) and should thus be important structural knowledge that is functionally preceding the acquisition of higher level speech processing abilities in humans like word or syllable segmentation.

The perceptual experiments with humans and animals, however, revealed high variability in their methodology. Only three studies have thus far experimented with a rhythm-only condition, i.e., *flat sasasa* (Ramus et al., 2003; Ramus & Mehler, 1999; White et al., 2012). Ramus et al. (2000) worked with normal speech as well as *saltanaj* speech (see Section 1); Tincoff et al. (2005)

worked with non-manipulated speech; Toro et al. (2003) again worked with *saltanaj*; Ramus (2002) has worked with non-manipulated speech, *saltanaj*, *saltanaj* + artificial intonation and *flat sasasa* + artificial intonation. A more recent study by White et al. (2012) used *flat sasasa* (this study is further discussed below). As a result of these methods, it is apparent that in almost all experiments intonation was actually present; however, in the interpretation of the data it was attributed little importance. This was possibly due to the fact that rhythmic durational cues have been taken for granted to be responsible for listeners' language discrimination ability for which the results from Ramus and Mehler (1999) provided strong evidence. The way the intonation contour was treated in that study, however, has already been criticized in the introduction to the present paper. The intonation-only condition (*aaaa*) Ramus and Mehler might have used highly distorted intonation contour information. Ramus and Mehler argued that intonation might be a necessary cue to distinguish between languages that belong to the same rhythm class, e.g., Spanish and Italian (p. 517), while it would be unnecessary and possibly even irritating for the discrimination of languages that belong to different rhythm classes.

In summary, the contribution of intonation to listeners' language discrimination ability is still unclear. The aim of the present experiment was to take a between-rhythm class contrast (English-French) and within-rhythm class contrast (French-Spanish) and test whether listeners' inability to distinguish between these language pairs might be enhanced when presented with *sasasa* that contains the intonation contour from the speech signal it derived from. Given the argumentation in Ramus and Mehler (1999), we should predict that listeners' accuracy for the French-Spanish contrast should be enhanced but for the English-French contrast intonation should further distort the durational cues and deteriorate listeners' discrimination performance. However, given the fact that performance was not above chance level for this group in the previous experiment based on *flat sasasa* speech, this effect is impossible to obtain.

4.1. Method

4.1.1. Subjects

24 native Swiss German speakers who were either students or had completed their tertiary education participated in this experiment, for which they were reimbursed. The listener group was randomly split into two subgroups of 12 listeners each, one taking part in the experiment with French and English, the other one in the experiment with French and Spanish.

4.1.2. Stimuli

The exact same sentences that were used for the English-French condition in Experiment II were used in the present experiment. Intonation was manipulated in the following way: the pitch contour was extracted using Praat's functions "To Pitch..." and "Down to Pitch tier." The "Pitch tier" object was then used to replace the pitch in the *flat sasasa* stimuli to create *sasasa* stimuli with the original intonation contour. Since the average pitch and its variability carries speaker specific information that listeners might use in the test session, we normalized average pitch by setting each pitch contour to a typical female pitch of 200 Hz and the coefficient of variation to 20% for each signal (i.e., a standard deviation of +-40 Hz was applied). The resulting *sasasa* signals were called "*intonation sasasa*."

4.1.3. Procedure

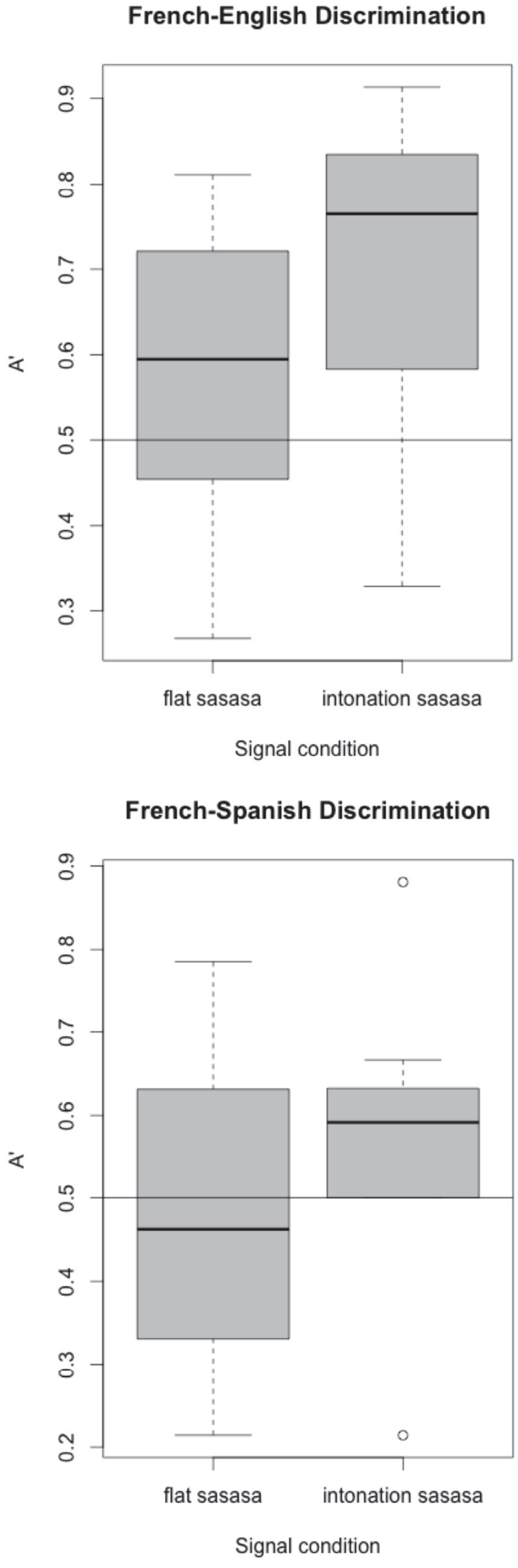
Listeners were tested according to the same procedure as in Experiment I.

4.2. Results

Figure 2, top box-plot, shows the distributions for listener sensitivity for the *flat sasasa* from Experiment I and the *intonation sasasa* for the French-English contrast. For the *intonation sasasa* listeners performed notably better than for the *flat sasasa*. A one-tailed *t*-test was carried out to test whether performance was higher than chance (Bonferroni corrected alpha: 0.025; 0.05/2 condition). The effect turned out significant for the intonation condition ($t[11] = 3.28$; $p = 0.0037$).

Figure 2, bottom box-plot, shows listeners performance for *intonation sasasa* for the French-Spanish contrast in comparison to the French-Spanish *flat sasasa* from Experiment II. It is visible that with *intonation sasasa* a considerably smaller part of the distribution is in the negative *A'* condition but nevertheless the bulk of the distribution is only just above chance level. A one-tailed *t*-test to test whether the performance is above chance was insignificant ($t[11] = 1.68$; $p = 0.06$). The effect and the descriptive impression, however, show a slight tendency to drift towards the *A'* area above 0.5. Like in Experiment II, it is possible that with some more participants an effect might have been obtained. But again, the magnitude of such an effect must be expected to be very weak. Given the outlier at about 0.2 *A'* it is also possible that more listeners would move the distribution below 0.5. The example shows very nicely that listener performance can vary tremendously for this task. While one listener has a performance close to perfection (*A'* about 0.9) the before mentioned listener is at the exact opposite (*A'* about 0.2). It is possible that individual listeners reach extremely high or low performance values by chance; but it is also possible that certain listeners are better at such tasks than others because they might pay more or less attention to either duration or intonation cues. It would be interesting to test the listeners with extreme *A'* value in more detail in future experiments.

Figure 2: Box-plots showing the distribution of listener accuracy for *flat sasasa* from Experiment I and for *intonation sasasa* for French-English (top) and French-Spanish discrimination (bottom).



4.3. Discussion

What are the cues that listeners pay attention to when discriminating French from English based on prosodic characteristics? The results from the present experiment clearly suggest that intonation plays an important role. This is the only condition in which listeners performed significantly above chance for the French-English discrimination. And this result is more than surprising. In both Ramus and Mehler's (1999) *flat sasasa* and *aaaa* condition and in our intonation condition, participants' language discrimination ability for two languages that do not belong to the same rhythm class was tested (mora- vs. stress-timed languages in Ramus and Mehler; stress- vs. syllable -timing in our case). In Ramus and Mehler the performance went down when only intonation was presented; in our constellation, performance went up when intonation was added. This suggests that Ramus and Mehler's view regarding the confusing influence the intonation contour might have on participants' perception is not supported by our data. On the contrary: while our participants were not able to distinguish between English and French based only on speech rhythm, they performed well when intonation was added. As such, our data provided the first evidence that the discrimination of languages between rhythm classes is not per se dependent on durational variability between these languages. Further, Ramus and Mehler's hypothesis that languages of the same rhythm class which cannot be distinguished purely on the basis of speech rhythm could be better discriminated when intonation is added as an extra cue, cannot be confirmed through this experiment either: for the French-Spanish group we did not obtain a difference in performance when intonation was added.

5. GENERAL DISCUSSION AND CONCLUSIONS

In the present research we demonstrated that Swiss listeners were able to distinguish between English and Japanese based on durational cues to consonantal and vocalic intervals only (Experiment I). This replicated previous findings for French listeners by Ramus and Mehler (1999). In principle this result also confirmed widely held beliefs that listeners are able to discriminate languages from different but not the same rhythm class based on prosodic cues (class distinction hypothesis). The results from Experiment II, however, provided strong counter evidence. According to the theory we predicted that discrimination between languages is possible when stress- and syllable-timed languages are paired, as in English-French, or syllable- and mora-timed languages, as in French-Japanese and Spanish-Japanese (the rhythm class affiliation of Spanish, however, was argued to be unclear). Our results, however, revealed no discrimination accuracy above chance for these language contrasts. This was also true for the language contrast that was described as one of the most canonical stress-/syllable-timed contrast, i.e., English and French. We must therefore conclude that discrimination between languages of different rhythm classes is not always possible based on durational cues. In Experiment III we added intonation to the previously not solvable discrimination task English-French and we found that listeners then showed discrimination ability that was well and significantly above chance. We did not find this effect for the Spanish-French within-class contrasts. From Experiment III we must conclude that previous hypotheses about the role of intonation in listeners' language discrimination ability have to be revised. According to our results it seems wrong to assume that intonation may distort listeners' ability to discriminate between languages from the different rhythm classes and it further seems wrong to assume that intonation might provide necessary cues to discriminate between languages of the same rhythm class. This means that intonation might play a much more important role in listeners' language discrimination ability than widely assumed based on the rhythm class hypothesis.

What do these results tell us about a possible effect of rhythm class? We see two ways of interpretation: (a) For the researcher who wishes to hold on to the rhythm class hypothesis the results might mean that some between-rhythm class contrasts, like French and English, rely on intonation. In other words, one could hypothesize that the only reason why we hear a more regular timing of syllables in French is because possible psycho-acoustic effects triggered by syllabic intonation movements are a possible reason for the sensation of durational regularity in the signal. Without these intonation movements the perceptual durational regularity would not be present. Support for this theory may be seen in the results for the French-Spanish within-rhythm class contrast, where even addition of intonation did precisely not lead to an improvement in performance. (b) The opponents of the rhythm class hypothesis might argue that the results do not provide any evidence in support of the class distinction hypothesis. One might argue that our languages show a wide variety of characteristics regarding their prosodic structure, which may or may not vary between individual language contrasts, always dependent on which particular languages are being paired. In our case it happened that between English and French intonation is more salient, whereas between English and Japanese durational characteristics are more salient. Between French and Japanese it was not the durational characteristics but possibly intonation. No test results are available for this condition.

At this point we leave it to the reader which point of view he/she wishes to take: the rhythm class supporting or the rhythm class opposing point of view. To answer this question unambiguously, a further study should be conducted in which, for example, more within class contrasts in the presence of intonation would be tested. Should within-class contrasts be found in which listeners' performance is enhanced by the presence of intonation, it would mean that the pro-rhythm class view cannot be held any longer.

6. ACKNOWLEDGEMENTS

We wish to thank Franck Ramus for sharing his speech recordings used in Ramus and Mehler (1999) and Ramus et al. (1999) and which were used for the present experiments. The work by the second author was supported by the Swiss National Science Foundation (Grant number: 100015_135287) and the Gebert-Rüf Stiftung (GRS-027/13).

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics.* Chicago, IL: Aldine.

Almeida, M. (1997). Organización temporal del español: el principio de isocronía. *Revista de Filología Románica, 14*(1), 29-40.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*, 351-373. http://dx.doi.org/10.1016/j.wocn.2012.02.003

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 15-Barcelona, Spain*), 2693-2696.

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

Dancovičová, J., & Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 16-Saarbrücken, Germany),* 1241-1244.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*(1), 51-62.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), *Language and language processing* (pp. 231-241). Frankfurt am Main, Germany: Peter Lang.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In N. Warner & C. Gussenhoven (Eds.). *Papers in laboratory phonology 7* (pp. 515-546). Mouton de Gruyter. http://dx.doi.org/10.1515/9783110197105.2.515

Ladefoged, P. (1967). *Linguistic phonetics*. Los Angeles: Phonetics Laboratory, University of California.

Ladefoged, P. (1975). *A course in phonetics*. New York, NY: Harcourt Brace Jovanovich.

Leemann, A., Dellwo, V., Kolly, M.-J., & Schmid, S. (2012). Rhythmic variability in Swiss German dialects. In *Proceedings of the 6th International Conference on Speech Prosody*, 607-610. Shanghai, PRC.

Lloyd James, A. (1929). *Historical introduction to French phonetics*. London, UK: ULP.

Mairano, P. (2011). *Rhythm typology: Acoustic and perceptive studies* (Doctoral dissertation). University of Turin, Italy.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756–766.

O'Connor, J.D. (1965). The perception of time intervals. *UCL Working Papers in Phonetics and Linguistics, 2*, 10-15.

Ohala, J., Riordan, C. J., & Kawasaki, H. (1979). Investigation of pulmonic activity in speech. In *Proceedings of the 9th International Congress of Phonetic Sciences (ICPhS 9-Copenhagen, Denmark*), 205.

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

Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, *8*(3), 293–304.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual*

Review of Language Acquisition, 2(1), 85-115. http://dx.doi.org/10.1075/arla.2.05ram

Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 15-Barcelona, Spain*), 337-342.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*, 349-351. http://dx.doi.org/10.1126/science.288.5464.349

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America, 105*(1), 512-521.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265-292. http://dx.doi.org/10.1016/S0010-0277(00)00101-3

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic controversies* (pp. 73-79). London, UK: Edward Arnold.

Shen, Y., & Peterson, G.G. (1962). *Isochronism in English*. Buffalo, NY: Department of Anthropology and Linguistics, University of Buffalo.

Tincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F., & Mehler, J. (2005). The role of speech rhythm in language discrimination: Further tests with a non-human primate. *Developmental Science, 8*(1), 26–35. http://dx.doi.org/10.1111/j.1467-7687.2005.00390.x

Toledo, G.A. (1988). *El ritmo en el español.* Madrid, Spain: Gredos.

Toro, J. M., Trobalon, J. B., & Sebastián-Gallés, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition, 6*, 131-136. http://dx.doi.org/10.1007/s10071-003-0172-0

White, L., & Mattys, S.L. (2007). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró & M.-J.Solé (Eds.), *Segmental and prosodic issues in Romance phonology. Current issues in linguistic theory series* (pp. 237-257). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/cilt.282.16whi

White, L., Mattys, S.L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*, 665-679. http://dx.doi.org/10.1016/j.jml.2011.12.010