Amalia Arvaniti* (University of California, San Diego)

Rhythm classes and speech perception

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 not strongly supported 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 have not found 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 – which is said to reflect differences in syllable complexity and vowel reduction between stress- and syllable-timed languages – and to use the related scores in order to place a language in one of the rhythm classes.

The initial success of metrics has boosted the previously waning support for rhythm classes. Further, it has provided the impetus for proposals that highlight the important function of rhythm in language acquisition and speech processing. Specifically, it has been suggested that infants utilize information about the extent and variability of vocalic intervals in their native language to

* I thank the members of the Institute for Language and Speech Processing (ILSP) in Athens who generously contributed their time, and in particular Spyros Raptis and his team for preparing the stimuli of Experiment B; their contribution cannot be understimated and I am greatly indebted to them. Special thanks are due to Marianna Katsoyannou who arranged my visit to ILSP, to Tristie Ross who prepared and ran most of Experiment A and to Tara Rodriquez who prepared and ran Experiment B and made many helpful suggestions for its design; thanks are also due to Julian Bauman, Sarah Choi, Noah Girgis, Christina Lee, Jini Shim and Amanda Simons for their help. I finally thank Oliver Niebuhr for his helpful comments at Speech Prosody 2010 and for inviting me to contribute to this volume, and Jennifer Cole, Chilin Shih, José Ignacio Hualde, Suzanna Fagyal and their students at the University of Illinois at Urbana-Champaign for valuable feedback. The financial support of the University of California, San Diego Committee on Research through grants LIN210G and LIN199G is hereby gratefully acknowledged.

deduce its rhythmic properties which, in turn, provide the basis for coarse-grained segmentation of the speech signal into stress feet, syllables or moras (among many, Nazzi, Bertoncini & Mehler 1998; Ramus & Mehler 1999; Nazzi, Jusczyk & Johnson 2000). Similarly, Cutler and her colleagues have shown that processing relies on the same prosodic unit as that of a language's rhythm class (Cutler et al. 1986; Culter & Otake 1994; Murty, Otake & Cutler 2007). Given the importance attributed to rhythm, it is essential to accumulate as much evidence as possible about the accuracy of the rhythm typology to which the function of rhythm in processing and acquisition is crucially linked.

This is all the more so because despite their initial success, metrics – which are currently the main evidence in support of rhythm classes – have been shown for some time to be problematic as measures of rhythm class differences. Grabe & Low (2002), who computed both PVIs and %V-ΔC for their corpus of 18 languages, found that the two sets of metrics classified some languages differently; e.g. Thai was classified as syllable-timed by %V-ΔC, but as stress-timed by PVIs. Arvaniti (2009; to appear) tested %V-ΔC and PVIs using a large corpus of data from English, German, Italian, Spanish, Greek and Korean. She found that metric scores show great inter-speaker variability and depend on the materials used and the method by which speech is elicited. These results add to an increasing body of research which has cast doubt on the effectiveness of metrics in rhythm classification and, by extension, on the rhythm typology they support (e.g., Barry, Andreeva & Koreman 2009; Loukina et al. 2009; see Arvaniti 2009 for a review).

Despite this evidence, the idea of rhythm classes remains strong. Because of this unwavering popularity, it is worth considering alternative bases for it that do not rely on timing relations in production. An obvious basis could be sought in perception, since the "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 speech 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 are not heavily 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. in her study, 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 intonation but kept the timing patterns of the original sentences by substituting consonantal intervals with [s] and vocalic intervals with [a] and replacing the original F0 modulation with flat slightly declining F0, a manipulation known as *flat sasasa*. They found that *some* pairs of languages belonging 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. Other pairs of languages, however, did not show the same pattern. In particular, Ramus et al. (2003) found that Polish – which Ramus et al. (1999) had classified as stress-timed – is discriminated from both stress-timed English and syllable-timed Spanish.

The disagreements between these studies are likely due to a number of confounds. The tapping 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, they 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 as stress- or syllable-timed baffling, while Miller herself admits that phoneticians taking part in her experiment may have been influenced by their training. Finally, the Polish results of Ramus et al. (2003) that are not compatible with the idea of rhythm classes suggest that additional factors may aid language discrimination. This conclusion is supported by other studies showing discrimination between languages of the same rhythm class on the basis of degraded signals, such as Mofta & Roach (1988) on English vs. Arabic (for a review of discrimination and identification results showing that both are possible with degraded signals and independent of rhythm class, see Komatsu 2007).

In short, the existing results on the perception of speech rhythm are inconclusive and point to methodological weaknesses. Specifically, results like those of Miller (1984) suggest that listeners – especially those who are linguistically naïve – cannot satisfactorily deal with explicitly categorizing languages into rhythm classes. On the other hand, tapping may be too indirect and ultimately may not have to do with speech rhythm at all but, rather, with general properties of motor control. Finally, language discrimination as in the experiments of Ramus et al. (2003) could be due to a number of differences inadvertently present in the stimuli; e.g. in a series of experiments similar to those of Ramus et al. (2003), Rodriquez & Arvaniti (2011) found that language discrimination can be achieved based on tempo and F0 differences rather than rhythm class. These issues with the existing research point to the need for a different experimental paradigm that goes beyond simple discrimination and is neither as indirect as tapping nor as explicit as categorization.

The two experiments presented here use a novel paradigm the aim of which is to avoid the problems with the paradigms discussed above while relying on

the principles that underlie the differences between rhythm classes. Specifically, the difference between stress- and syllable-timing lies in their basic rhythmic patterns: syllable-timed languages have a rhythm akin to a simple cadence with every syllable being rhythmically of equal value to all the others in an utterance; stress-timed languages, on the other hand, have a rhythm closer to a series of trochees, with each stress foot beginning with a stressed syllable followed by a small number of unstressed syllables (indeed, metrical accounts of rhythm suggest that in stress-timed languages like English, a maximum of two unstressed syllables are tolerated in a foot and the presence of two unstressed syllables is licensed only in certain circumstances; Hayes 1995). As noted earlier, the exact phonetic manifestation of these differences in rhythmic patterning is not fully agreed upon: traditional views favor isochrony of syllables or stress feet, while more recent research favors differences in syllable structure and vowel reduction patterns. In either case, however, the difference in the percept of basic rhythmic pattern stands.

The present research paradigm exploits these differences: if the principled distinction between cadences and downbeat-initial meters characterizes syllable- and stress-timing respectively, then listeners should hear utterances from stress-timed languages as more similar to a series of trochees than utterances from syllable-timed languages. This hypothesis was tested by having listeners use a scale from 1 to 7 to rate the similarity between a sequence of non-speech trochees and stimuli from six languages belonging to different rhythm classes. The hypothesis was that stimuli from stress-timed languages would be rated more highly (i.e. more similar to the trochees series) than stimuli from syllable-timed languages. In turn, the different ratings would provide indirect evidence that listeners can rhythmically categorize languages into different classes.

Further, since Miller (1984) has shown that rhythm classification can be influenced by the native language of the listeners, the first experiment also explored the extent to which the listeners' ability to perform this task and their ratings of different languages depend on their native tongue. In particular, it was expected that English listeners would 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 very trochee-like (compared to the other two participant groups). The opposite was expected from the Greek participants, who speak a language with strong stresses but tolerate irregularities in prominence patterns much more than speakers of English do (Arvaniti 2007). Finally, Korean participants were expected to find the test more difficult than the other two groups since their native language lacks stress altogether (Jun 2005).

2 Experiment A: Method

2.1 Experiment A: Participants

Three groups of listeners took part in the study; 23 were native speakers of Standard Athenian Greek, 22 were native speakers of Southern California English, and 21 were native speakers of Seoul Korean. All participants were monolingual, though the Greek and Korean speakers also spoke English as L2. The average age was 36.5 for the Greeks, 20.75 for the Americans, and 23 for the Koreans. The Greek participants were recruited and tested in Athens, Greece and were all volunteers. The American and Korean participants were recruited from the student population of the University of California, San Diego (UCSD). They all participated for course credit, except for five Korean listeners who were graduate students and were paid a small fee. All participants were naïve as to the purposes of the experiment and reported no history of speech or hearing problems. None of the Korean participants had any training in linguistics in general or phonetics in particular. Two of the American participants (9% of the group) had some undergraduate training in linguistics, while six of the Greek participants (26% of the group) had a Master's or Ph.D. in linguistics, with two of these having a specialization in phonetics. None of the linguistically trained participants guessed the purpose of the experiment and their responses did not deviate in any respect from those of the other subjects; thus there is no reason to believe that these differerences in background have any bearing on the results.

2.2 Experiment A: Stimuli

The stimuli for Experiment A were sentences selected from a large corpus collected for the production study of Arvaniti (2009; to appear). The corpus included data from two stress-timed languages, English and German, two syllable-timed languages, Italian and Spanish, and two languages weakly classified as syllable-timed, Greek and Korean (see Arvaniti to appear, for a review of studies on Greek and Korean rhythm). In order to keep the experiment to a reasonable length, 12 sentences of each language were selected from the data of one male and one female talker of each language (i.e. there were six sentences per talker) for a total of 72 stimuli.

The two talkers of each language were chosen using fluency as the criterion for selection. The six sentences of each talker were chosen on the basis of two main criteria. First, the sentences had to be as similar as possible across talkers and languages, both in terms of number of syllables and in terms of overall duration. Second, the sentences of each talker were selected in equal numbers

from the three subsets of sentences in the corpus of Arvaniti (to appear). These subsets included "stress-timed" 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; for details, see Arvaniti to appear). The selected English sentences are shown in Table 1. The aim of choosing stimuli along these lines was to test whether the greater or lesser temporal variability of the "stressed-timed", "syllable-timed" and "uncontrolled" stimuli would have an effect on responses, for example by increasing the ratings of "stress-timed" stimuli as compared to "syllable-timed" ones. This temporal variability is reflected in %V and ΔC scores which were significantly lower for "syllable-timed" stimuli than "stress-timed" and "uncontrolled" stimuli [for %V, $F(2,36) = 7.3$, $p < 0.002$; for ΔC, $F(2,36) = 20.4$, $p < 0.0001$].

In order to mask the language of the stimuli, 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, the amplitude of the stimuli was adjusted 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. The "frog" sound consists of two rapidly decaying irregular pulses connected by similar pulses of much lower amplitude. Each of the louder pulses of the "frog" is 7 ms in duration with the first pulse being higher in amplitude; energy ranges between 1100 Hz and 2900 Hz. The overall auditory percept is of a single sound that resembles a cross between the tick of a clock and a chirp. In order to make the sequence of two "frogs" sound like a trochee, the second repetition was made shorter and softer than the first: the first repetition was 220 ms long while the second one was 120 ms; the mean intensity of the loudest part of each "frog" (i.e. of the first pulse) was 70 dB for the first repetition and 66 dB for the second (see Figure 1). 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 the aim was for participants to focus their comparison on the *rhythm* rather than the *tempo* of the trochees and the stimuli. The effective tempo of the trochee series was 3 "frogs" per second, a rate different from those of the linguistic stimuli (see Table 2).[1]

[1] Tempo was calculated by dividing the number of vocalic intervals in an utterance by its duration including any pauses. Although this calculation understimates tempo in actual speech, it more accurately reflects the tempo of the stimuli, in which the distinctions between vowels in hiatus could not be heard and pauses could not be clearly distinguished from stop closures.

Sentence class	Sentences used as stimuli
"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.*

Table 1: The English sentences used as stimuli.

The stimuli were collated using PRAAT to create one sound file in which each trochee-and-stimulus sequence was heard once for a total of 72 trials. The experiment lasted approximately 12 minutes. Two randomization orders were prepared and counterbalanced across participants within each language group. There was 1 s of silence between the trochee series and the following linguistic stimulus, and 2 s of silence between trials (see Figure 1). In all trials, the trochee sequence preceded the linguistic stimulus since (a) the aim was to compare the linguistic stimuli to the trochees and (b) switching this order would make harder an already difficult task and would double the duration of the experiment.

	Experiment A			Experiment B		
	female talker	male talker	language mean	female talker	male talker	language mean
English	5.5 (0.6)	5.0 (0.2)	5.3 (0.5)	5.6 (0.7)	4.9 (0.4)	5.3 (0.7)
German	4.5 (0.4)	5.8 (0.6)	5.1 (0.8)	4.5 (0.4)	5.8 (0.8)	5.2 (0.9)
Greek	6.2 (0.6)	8.0 (0.8)	7.1 (1.2)	6.0 (0.7)	7.0 (0.5)	6.5 (0.8)
Italian	6.7 (0.4)	6.4 (0.4)	6.2 (0.4)	6.6 (0.4)	6.1 (0.6)	6.3 (0.5)
Korean	5.8 (0.4)	6.5 (0.2)	6.1 (0.5)	5.5. (0.3)	6.3 (0.5)	5.9 (0.6)
Spanish	5.9 (0.4)	5.5 (0.5)	5.7 (0.5)	5.4 (0.4)	5.5 (0.6)	5.4 (0.5)

Table 2: Mean tempi and standard deviations (in brackets) of the stimuli of Experiments A and B; data presented pooled for each language, as well as separately for each talker.

2.3 Experiment A: Procedures

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

The participants were provided with an answer sheet which 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 Likert

scale of 1 (most dissimilar) to 7 (most similar) was used for rating. 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 randomly selected from the data of the same talkers.

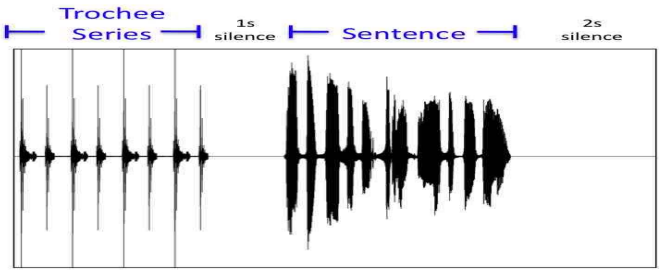

Figure 1: Timeline of a trial in Experiment A.

Many participants reported that they found the task difficult but the vast majority completed it without problems. Two English and two Greek questionnaires were discarded because participants did not provide responses to several stimuli. One Greek participant did not complete the experiment because she was called away, and the questionnaire of another was discarded after preliminary analysis (Arvaniti & Ross 2010), as it was noted that she utilized exclusively one end of the scale. Finally, two Korean participants were excluded because they disclosed after completing the experiment that they had arrived in the US well before adolescence and were thus sequential bilinguals of Korean and English. This brought the number of usable responses to 20 for English, 19 for Korean and 19 for Greek, for a total of 58.

3 Experiment A: Results

Kruskal-Wallis Analysis of Variance was used first to determine whether the responses of the three groups of listeners differed from each other. The results showed that the three groups responded to the stimuli in the same manner [$H(2, N = 4176) = 3.78$, n.s.]. Since the responses did not show an effect of participant language, they were pooled for subsequent analyses.

Specifically, ordinal probit regression was run, using ratings as the dependent variable and stimulus language, stimulus rhythm class ("stress-timed", "syllable-timed", "uncontrolled") and talker gender as predictors. Talker gender was included in the analysis, since (impressionistically) the female talkers spoke more carefully than the male talkers, and this difference could have an effect on responses.

The analysis showed was that stimulus language and stimulus class did affect responses [for language, Wald = 17.7, $p < 0.003$; for stimulus class, Wald = 10.2, $p < 0.006$]. Wilcoxon matched pair tests showed that English was rated less trochee-like than all the other languages [$p < 0.009$ for all pairwise comparisons; see Table 3 for descriptive statistics]; no differences were found between the other five languages. In addition, "uncontrolled" stimuli were generally rated higher than the other stimuli, a difference that was statistically significant for the comparison between "uncontrolled" and "syllable-timed" stimuli [$p < 0.002$].

The effect of stimulus class was not the same across languages, as indicated by the interaction between language and stimulus class [Wald = 28.1, $p < 0.002$]. Stimulus class affected the responses to Greek and Korean stimuli (see Table 4 for descriptive statistics): in Greek, "uncontrolled" stimuli were rated significantly higher than both "stress-timed" and "syllable-timed" stimuli [$p < 0.02$], while in Korean "syllable-timed" stimuli were rated lower than "stress-timed" and "uncontrolled" stimuli [$p < 0.007$].

In addition, the analysis showed an interaction between talker gender and language [Wald = 11.4, $p < 0.04$]. Talker gender affected responses in the Greek and Italian stimuli: for Greek, the stimuli of the female talker received higher ratings than the stimuli of the male talker, while the opposite obtained in Italian [$p < 0.02$ in both cases; see Table 5 for descriptive statistics].

		Median	Mode	Freq. of Mode
Exp. A	English	3	2	172
	German *	3	3	164
	Greek *	3	3	145
	Italian *	3	3	140
	Korean *	3	3	152
	Spanish *	3	3	146
Exp. B	English *	4	5	45
	German*	3	2	43
	Greek	3	2	43
	Italian	3	2	57
	Korean	3	2	50
	Spanish *	4	2	47
	Controls *	6	7	76

Table 3: Descriptive statistics of responses pooled by stimulus language in Experiments A and B. Stars indicate languages that were rated significantly higher than the unstarred languages within each experiment (for details see §3 and §6).

4 Experiment A: Interim Discussion

Overall, the results did not provide evidence in favor of rhythm classes. All three groups of participants unanimously rated English less trochee-like and grouped it separately from the other languages, contrary to what one would expect if ratings were based on rhythm class (in which case, English should

have been grouped with German). Finally, the ratings suggest that differences in the syllable complexity of the stimuli do not by and large make listeners perceive stimuli with greater temporal variability as more trochee-like than others (if that had been the case, "stress-timed" stimuli would have had higher ratings than "syllable-timed" ones).

Although the above results provide prima facie evidence that this type of indirect categorization is not easy for listeners, it is possible that the results were due, at least in part, to additional reasons. First, it is possible that the simple trochee pattern was not sufficiently close to the rhythm of any of the languages (a possibility that in itself suggests that speech rhythm is more complex than timing; cf. Arvaniti 2009; Rodriquez & Arvaniti 2011). Second, the difficulties that participants reported could be related to the nature of the stimuli. Specifically, despite warnings that comprehension would not be possible, many participants mentioned that they were trying to guess what languages they were listening to and what was being said; others mentioned that they found the muffling that low-pass filtering created irritating. These reactions suggest that many participants may have been distracted by the experimental setup. Experiment B aimed at addressing these problems by using flat *sasasa* a degraded signal transform that retains only (some) temporal information (see §1). This change was expected to have two effects: first, allow listeners to focus more on the task at hand, and second, make the stimuli more comparable to the trochees by stripping them of their complexity.

		Median	Mode	Mode frequency
English	"stress-timed"	3	2	56
	"syllable-timed"	3	2	56
	"uncontrolled"	3	2	60
German	"stress-timed"	3	3	57
	"syllable-timed"	3	3	58
	"uncontrolled"		3, 4	49, 49
Greek	"stress-timed"	3	3	61
	"syllable-timed"	3	3	53
	"uncontrolled" *	4	5	49
Italian	"stress-timed"	3	3	43
	"syllable-timed"	3	2	53
	"uncontrolled"	4	4	51
Korean	"stress-timed" *	4	4	49
	"syllable-timed"	3	3	55
	"uncontrolled" *	3	3	55
Spanish	"stress-timed"	3	3	56
	"syllable-timed"	4	3	49
	"uncontrolled"	3	3	54
Pooled	"stress-timed"	3	3	312
	"syllable-timed"	3	3	278
	"uncontrolled"	3	4	301

Table 4: Descriptive statistics of responses by stimulus language and stimulus rhythm class in Experiment A. Stars indicate stimuli that were rated significantly higher than the unstarred stimuli of the same language (for details see §3).

		Median	Mode	Mode frequency
English	female talker	3	2	94
	male talker	3	2	78
German	female talker	3	3	81
	male talker	3	3	83
Greek	female talker *	3	3	73
	male talker	3	2	75
Italian	female talker	3	3	78
	male talker *	4	2	67
Korean	female talker	4	4	71
	male talker	3	3	83
Spanish	female talker	3	3	72
	male talker	3	3	74

Table 5: Descriptive statistics of responses by stimulus language and talker gender in Experiment A. Stars indicate stimuli that were rated significantly higher than the unstarred stimuli of the same language (for details see §3).

5 Experiment B: Method

Seventeen listeners took part in Experiment B, 13 females and 4 males. Ten of them were monolingual speakers of English and the other seven were bilinguals with English as their dominant language (with the exception of one subject who was a Spanish-German bilingual). All participants were recruited from the UCSD undergraduate community and took part for course credit. They were all in their early twenties. They were naïve as to the purposes of the experiment and reported no history of speech or hearing problems.

The same basic method used for Experiment A was employed for Experiment B, except that instead of being low-pass filtered the stimuli were converted into flat *sasasa*. In order to make the two experiments as comparable as possible except in terms of the signal tranform, data from the same talkers as before were selected. For the female talkers the same utterances were used as in Experiment A (with the exception of one of the English female talker's stimuli which was inadvertently replaced with a different sentence). For the male talkers, different utterances were selected – following the same criteria as in Experiment A – since flat *sasasa* could render the utterances of the male and female talker of each language too similar to each other. ANOVAs with experiment as the between-subjects factor and sentence duration, tempo, %V, ΔC, and PVIs as the dependent variables did not reveal any significant differences, suggesting that the stimuli of the two experiments were comparable except in the type of signal manipulation used.

In order to create the *sasasa* stimuli, [sa] diphones were extracted from the stressed syllable of the Greek name [ˈsasa] elicited within an utterance from one male and one female Greek speaker. On the basis of these diphones, stimuli were synthesized using the facilities of STRAIGHT in Matlab (http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html; I am

grateful to Spyros Raptis, ILSP, for synthesizing the stimuli). For each *sasasa* stimulus, the [s] and [a] durations were those of the consonantal and vocalic intervals respectively of the original sound file. F0 was slightly declining and spanned the middle third of the original talkers' range in the utterances from which the [sa] diphones were excised (97-120 Hz for the male talker and 220-268 Hz for the female talker). In addition, the *sasasa* stimuli differed from the low-pass filtered stimuli in two respects: first, pauses – found in five of the 72 stimuli – were not included in the *sasasa* versions; in two stimuli this resulted in consecutive intervals of the same type (consonantal in one, vocalic in the other) which were synthesized as one interval. Further, in order to ensure auditory uniformity of the stimuli, utterance initial vowels were excluded from synthesis so that all stimuli began with [s].

Some additional modifications to the original experiment were also made to facilitate the participants' task as much as possible. First, controls that were auditorily as close as possible to the trochee series were created by concatenating several copies of the original [ˈsasa] utterances. In order for controls not to sound identical, their length ranged from nine to fourteen syllables and their tempo varied by 10-15% from the original [ˈsasa]. In addition, listeners were given slightly longer to respond to each stimulus (2.5 s instead of 2 s), and a longer ISI of 3.5 s after every block of 14 trials; at the end of each block participants also heard a recorded warning that they should turn the page of their questionnaire.

6 Experiment B: Results

Kruskal-Wallis Analysis of variance was used first to test whether the linguistic background of participants (being monolingual or bilingual) had affected their responses; the results showed that this was not the case [$H(1, N = 1428) < 1$], so for all other analyses responses were pooled across subjects. Next, ordinal probit regression was used to determine whether the control stimuli were rated more similar to trochees than the other classes of stimuli. The analysis showed that stimulus class ("stress-timed", "syllable-timed", "uncontrolled" or control) did affect responses [$W = 158.6$, $p < 0.001$]. According to Wilcoxon matched pair tests, control stimuli were rated significantly higher than all other classes [$p < 0.0001$ for all pairwise comparisons], indicating that the listeners correctly interpreted and responded to the task (see Table 3).

Since the main aim of using the controls was to facilitate the participants' task they were included in further analyses. Ordinal probit regression not including the controls was run with the same three predictors as in Experiment A, stimulus language, stimulus rhythm class and talker gender. Results showed that once the controls were removed, stimulus language was the only predictor

[W = 50.1, $p < 0.0001$; stimulus class and talker gender were not significant; W = 0.04 and W = 0.02 respectively]. Wilcoxon tests showed that overall, English, German and Spanish were rated similarly to each other, as were Greek, Korean and Italian, and that the languages in the former group were rated more trochee-like than the languages in the latter (with the exception of German which was rated higher than Greek and Italian only). These results are presented in Table 3.

Finally, talker gender interacted with language [W = 27.8, $p < 0.0001$]. Wilcoxon matched pair tests showed that the stimuli of female talkers were rated significantly higher than those of male talkers in German [$p < 0.001$] and Korean [$p < 0.01$], while the opposite obtained in English [$p < 0.005$]. These results are presented in Table 6.

		Median	Mode	Mode frequency
English	female talker	4	2, 4	21, 21
	male talker *	5	5	26
German	female talker *	4	5	25
	male talker	3	2	24
Greek	female talker	4	5	25
	male talker	2.5	1	28
Italian	female talker	2.5	2	28
	male talker	3	2	29
Korean	female talker *	3	2, 3	23, 23
	male talker	3	2	27
Spanish	female talker	3	2	27
	male talker	4	4	21

Table 6: Descriptive statistics of responses by stimulus language and talker gender in Experiment B. Stars indicate stimuli that were rated significantly higher than the other stimuli of the same language (for details see §6).

7 Discussion and conclusions

The two experiments presented here showed that listeners can indirectly classify languages into groups by rating the similarity of their rhythm to that of non-speech trochees. However, the classification did not follow rhythm class lines except very weakly: in Experiment B, English and German were similarly highly rated, as expected, but so was Spanish. Thus, the results provide little support for rhythm classes.

In addition, despite the very similar set up and stimuli between the two experiments, the languages were classified differently in each. These differences question the idea that the overall impression of a language's rhythmicity improves when all aspects of prosody except the relative timing of consonants and vowels are stripped from the signal. If that were the case, then one would expect both experiments to yield very similar results, with

Experiment B just showing stronger effects. The different ratings in the two experiments point to interactions between components of prosody present in Experiment A but not in Experiment B. Specifically, the low-pass filtering used in Experiment A preserved both timing and intonational characteristics of the languages, while flat *sasasa* preserved only the former (in a simplified form). Clearly, the more faithful representation of prosody in Experiment A gave a different impression of rhythmicity. This conclusion is supported by Kohler (2008) who found that F0 information can significantly affect the percept of prominence, and those of Rodriquez & Arvaniti (2011) who show that F0 can influence language discrimination. Results like these indicate that timing information cannot be processed totally independently of other prosodic phenomena such as intonation; thus asking listeners to focus on timing alone may give an inaccurate idea of how rhythm is processed in real speech.

Similar conclusions are drawn if one considers the differences between the experiments regarding the rhythm class of the stimuli: while in Experiment A "uncontrolled" stimuli were overall rated more trochee-like than the rest, in Experiment B, no effect of stimulus class was observed. Again, if timing differences drive rhythmic impressions, one would expect Experiment B to show a strong effect of stimulus class on ratings since, in this experiment, listeners had only timing on which to base their judgments and timing was the a priori difference between classes of stimuli. The fact that Experiment B did not show any effect of stimulus rhythm class suggests that quite likely it was other prosodic features that gave rise to the related result in Experiment A. Finally, both experiments showed an effect of talker gender though the results again did not agree between experiments even though the stimuli came from the same talkers and largely the same utterances. This again supports the view that the *sasasa* transform does not give the same rhythmic impression as a signal in which prosody remains largely intact.

Given the above, it seems plausible that the different ratings in the two experiments were due to other prosodic features or even, in the case of talker-specific effects, differences in voice quality that were present in Experiment A but eliminated in Experiment B.

An obvious prosodic feature that could have influenced the results is speech tempo which has been shown before to interact with rhythm. For instance, Dellwo & Wagner (2003) and Russo & Barry (2008) have found that tempo affects variability in speech (as measured by rhythm metrics) while Loukina et al. (2009) have found that it significantly enhances the ability of metrics to discriminate between languages using computational methods.

Statistical analysis of the tempo of the present stimuli provides some support that tempo may have influenced responses, with slower tempi generally leading to higher ratings. As can be seen in Table 2, the Greek female talker and the Italian male talker who were rated more highly than the other talker of the same language in Experiment A both spoke more slowly and

with the same average tempo of 6.2 sylls/s (though the difference between the two talkers of each language is significant only for Greek; $p < 0.0006$ according to Scheffé tests). Similarly, in Experiment B, the German, Korean and English talkers who were rated more highly then the other talker of the same language were the slow talkers though their speaking rates were not significantly different from those of their faster counterparts or as comparable to each other (4.9 sylls/s for the English male talker, 5.5 sylls/s for the Korean female talker and 4.5 sylls/s for the German female talker). Table 2 also shows that English, German and Spanish, which were rated more highly than the other languages in Experiment B, had low average tempi compared to Greek, Italian and Korean which were spoken at significantly faster rates [$F(1,66) = 37.2$, $p < 0.0001$]. Similarly, the "uncontrolled" stimuli that were rated more highly than the "syllable-timed" stimuli in Experiment A had a significantly lower tempo overall (5.9 sylls/s as opposed to 6.3 sylls/s for the "syllable-timed" stimuli, a statistically significant difference; $p < 0.03$, according to Scheffé tests). On the other hand, no tempo differences were present across "uncontrolled", "stress-timed" and "syllable-timed" stimuli in Experiment B in which stimulus class did not affect ratings. Although these results support the idea of tempo as an important factor in the percept of rhythmicity, it is important to also note that in Experiment A, English and German had practically identical tempi but English was rated significantly lower than German. When considered all together then these results suggest that although tempo most likely played a part, it was only one of several factors affecting listeners' responses.

By and large these results agree with those of previous studies. Like other research paradigms, the paradigm used here showed that categorizing languages for rhythm class is not an easy task. The results also indicate that the role of the listeners' first language may be task-dependent since it is found with some tasks but not others. Finally, they suggest that prosodic properties other than timing may influence responses. This was most strongly shown here for tempo, a factor that may have also played a part in Ramus et al. (2003), in which the tempo of the Polish stimuli that were discriminated from both English and Spanish was inadvertently increased during conversion to *sasasa*.

In conclusion, the experiments presented here show that language classification into rhythmic classes cannot be easily achieved on the basis of listener impressions. Further, a comparison of the two experiments shows that responses to rhythm depend on the nature of the information provided to listeners. Yet, given that timing, tempo and intonation are always present in speech and are most likely integrated in perception, the idea that rhythm is based solely on timing seems unlikely. Thus, these data, like those of previous studies, cast doubt on the idea of distinct rhythm classes which rest exclusively on the timing characteristics of each language. 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. Such research will no doubt help elucidate the function of speech rhythm in general and in language acquisition and speech processing in particular.

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. and T. Ross (2010): Rhythm classes and speech perception. Proceedings of Speech Prosody 2010, Chicago, 11-14 May 2010. Available at http://speechprosody2010.illinois.edu/.

Arvaniti, A. (to appear): The usefulness of metrics in the quantification of speech rhythm. Journal of Phonetics.

Barry, W., B. Andreeva, B. and J. Koreman (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., J. Mehler, D. Norris and J. Seguí (1986): The syllable's differing role in the segmentation of French and English. Journal of Memory and Language 25, 385-400.

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

Dellwo, V. and P. Wagner (2003): Relations between language rhythm and speech rate. Proceedings of the XVth ICPhS, 471-474, Barcelona, Spain.

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

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

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

Hayes, B. (1995): Metrical Stress Theory: Principles and Case Studies. Chicago: The University of Chicago Press.

Kohler, K. J. (2008): The perception of prominence patterns. Phonetica 65, 257-269.

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.

Komatsu, M. (2007): Reviewing human language identification. Lecture Notes in Computer Science 4441, 206-228.

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

Loukina, A., G. Kochanski, C. Shih, K. Keane and I. Watson (2009): Rhythm measures with language-independent segmentation. Proceedings of Interspeech 2009, Brighton, UK.

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

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

50-52.
Murty, L., T. Otake and A. Cutler (2007): Perceptual tests of rhythmic similarity: I. Mora rhythm. Language and Speech 50, 77-99.
Nazzi, T., J. Bertoncini and J. Mehler (1998): Language discrimination by newborns: Toward an understanding of the role of rhythm. Journal of Experimental Psychology 24, 756-766.
Nazzi, T., P. W. Jusczyk and E. K. Johnson (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., E. Dupoux and J. Mehler (2003): The psychological reality of rhythm class: Perceptual studies. Proceedings of the 15th ICPhS, 337-340,Barcelona, Spain.
Ramus, F. and J. Mehler (1999): Language identification with suprasegmental cues: A study based on speech resynthesis. Journal of the Acoustical Society of America 105, 512-521.
Ramus, F., M. Nespor and J. Mehler (1999): Correlates of linguistic rhythm in the speech signal. Cognition 73, 265-292.
Rodriquez, T. and A. Arvaniti (2011): Rhythm, tempo, and F0 in language discrimination. Journal of the Acoustical Society of America 130, 2567.
Russo, M. and W. J. Barry (2008): Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo. Proceedings of Speech Prosody 2008, 419-422, Campinas, Brazil.
Scott, D. R., S. D. Isard and B. de Boysson-Bardies (1985): Perceptual isochrony in English and in French. Journal of Phonetics 13, 155-162.

Index

English, German, Greek, Italian, Korean, Spanish, rhythm, rhythm perception, rhythm class, rhythm class discrimination, rhythm metrics, speech timing, tempo, language discrimination

Short Portraits

Arvaniti, Amalia

Associate Professor at the Department of Linguisitics, University of California, San Diego. Research interests: the phonetics and phonology of prosody, especially of speech rhythm and intonation, intonation and focus, sociophonetics, sociolinguistics, Greek, Romani.