The role of rhythm class, speaking rate and F_0 in language discrimination

Amalia Arvaniti[1,2*] & Tara Rodriquez[2]

[1]University of Kent, [2]University of California San Diego

Abstract

The division of languages into stress- syllable- and mora-timing is said to be supported by experiments showing that languages are discriminated only if they belong to different rhythm classes, a distinction said to be reflected in the duration and variability of consonantal and vocalic intervals (*timing*). The role of rhythm classes in discrimination is tested here along with the alternative that discrimination is due to speaking rate and F_0 differences and is independent of rhythm class. Five AAX experiments with English as *context* (AA) and Polish, Danish, Spanish, Greek or Korean as *test* (X) were conducted using the *sasasa* transform and modifying F_0 and speaking rate so as to compare responses to stimuli that retained the original speaking rate and F_0 of each language or were stripped of this information (speaking rate, F_0 or both) while retaining their timing. Discrimination was possible both across and within rhythm classes when speaking rates differed between context and test but largely impossible once speaking rate differences were eliminated; F_0 also played a significant if less consistent role in discrimination. The changes in responses associated with speaking rate and F_0 indicate that language discrimination arises from interactions between prosodic factors and that timing contributes but little. Consequently the results cast doubt both on the ecological validity of the *sasasa* transform, which brings timing to the fore while eliminating F_0 modulation, and on the rhythm class typology said to be reflected in timing distinctions.

[*] Corresponding author. Address for correspondence: Department of English Language and Linguistics, School of European Culture and Languages, University of Kent, Canterbury, Kent CT2 7NF, UK. Email: A.Arvaniti@kent.ac.uk

1. Introduction

A popular view of speech rhythm advocates that languages fall into distinct rhythm classes, stress- and syllable-timing (e.g. Abercrombie 1967), with mora-timing sometimes proposed as an additional category (e.g. Port, Dalby, and O’Dell 1987; Ramus, Nespor, and Mehler 1999). This typology is intricately linked with the notion of *timing*, a cover term typically referring to all aspects of durational variation in speech (cf. Arvaniti 2009), but used within the context of rhythm classes with two specific meanings. In the original conception of the rhythm class typology timing referred to isochrony, the claim that one prosodic unit in each class—the stress-foot, the syllable and the mora in stress-, syllable-and mora-timed languages respectively—has stable duration and it is its repetition at regular intervals that creates rhythm (e.g. Abercrombie 1967). More recently, a different aspect of timing has been seen as the cornerstone of rhythm class distinctions, namely the relative variability of vocalic and consonantal interval durations. Specifically, stress-timed languages are said to show greater variability in the duration of consonantal intervals because they allow for complex consonant clusters; in turn the presence of clusters reduces the percentage of speech taken up by vocalic intervals, while the presence of vowel reduction results in greater vocalic interval variability. Syllable-timed (and mora-timed) languages, on the other hand, are said to exhibit the opposite pattern, i.e. lesser variability of vocalic and consonantal interval duration and a greater proportion of speech made up of vocalic intervals (Ramus et al. 1999; Grabe and Low 2002). Despite their differences, both conceptualizations espouse the view that speech rhythm is based on the characteristics of specific temporal intervals, that is on some aspect of timing. As a result, speech rhythm research has focused on the measuring of durations in the acoustic signal and on the role of duration in perception, while the role of other components of prosody, such as F_0, has been largely overlooked. The present study examines the view of rhythm as a typology reflected in the relative variability of consonantal and vocalic interval durations (henceforth referred to as *timing*) by means of five perception experiments that test the role of timing, speaking rate and F_0 in between-language discrimination.

1.1. *Background and hypotheses*

Despite the enduring popularity of the rhythm class typology, evidence from production studies remains inconsistent at best. Evidence for isochrony has not been found in any language tested (see Bertinetto 1989, Arvaniti 2009, Kohler 2009, Fletcher 2010 for reviews). More recently, various methods (henceforth *timing-based metrics*) have been proposed to quantify the alternative view that rhythm class depends on the greater or lesser variability of consonantal and vocalic interval durations. The efficacy of several such metrics—%V and Δ*C* (Ramus et al. 1999), PVIs (Grabe and Low 2002) and Varcos (Dellwo 2006)[1]—at capturing rhythm class differences was tested by Arvaniti (2012a) using a large sample of data from English, German, Greek, Italian, Korean and Spanish. The data showed no clear separation of either languages or rhythm classes; rather, they revealed substantial variability across speakers, utterances (largely due to phonotactics), and elicitation styles (particularly between scripted and spontaneous speech). These results are further supported by Horton & Arvaniti (subm.) who found that the classification of languages into rhythm classes using naïve Bayes classifiers is inconsistent and depends on the exemplars of each class used to train the classifiers. In addition, both quantitative analysis (Payne et al. 2011; Renwick 2011) and unsupervised clustering (Horton & Arvaniti subm.) have confirmed that metrics are heavily influenced by utterance-specific phonotactics. All together the results of these studies show that values obtained using timing-based metrics cannot be seen as immutable properties of the languages involved but, rather, as points in wide distributions which overlap substantially across languages. In turn this substantial overlap strongly argues against the view that languages fall into rhythm classes with distinct timing characteristics (cf. Loukina et al. 2011 for similar conclusions).

The strongest evidence in favor of rhythm classes and the primacy of timing in determining rhythm class affiliation comes from perception research, particularly experiments that rely on discrimination. A series of such experiments show that discrimination between languages is possible if the languages involved belong to different rhythm classes (e.g. Nazzi, Bertoncini, and Mehler 1998; Ramus et al. 1999; Nazzi, Jusczyk, and Johnson 2000; Nazzi and Ramus 2003; Ramus, Dupoux, and Mehler 2003). Several of these experiments are based on variations of the *flat sasasa* transform in which consonantal intervals are replaced by [s] and vocalic intervals by

[a] while F_0 modulation is eliminated by replacing the F_0 curve with flat slightly declining F_0 (e.g. Ramus and Mehler 1999; Ramus et al. 2003). Since this transform encodes differences in the timing characteristics of vocalic and consonantal intervals and has been successful in showing discrimination, the results have been interpreted as support for the view that languages belong to distinct rhythm classes the differences among which are encoded in timing.

However, there are problems with this interpretation. First, discrimination experiments have largely been conducted using a small set of languages: Germanic languages (usually English, Dutch or German) have been used as representatives of stress-timing, Romance languages (usually Spanish, Catalan or Italian) as representative of syllable-timing, and Japanese as the sole representatives of mora-timing. However, since the languages within each group are closely related they share similarities that go beyond rhythm class. One of these similarities relates to speaking rate: as a group, the Germanic languages tested in the relevant literature are spoken at a relatively slow speaking rate of between 4.5 and 5.5 syllables/s. (e.g. Quené 2007, and references therein on Dutch; Clopper and Smiljanic 2011 on English; Pellegrino, Coupé and Marsico 2011 on German and English). The Romance languages usually tested are spoken at faster rates of above 6 syllables/s. (e.g. Dauer 1983, and references therein, on Spanish and Italian; Pellegrino et al. 2011 on French, Italian, and Spanish). Japanese is also spoken at a fast rate of 7.84 sylls/s. (Pellegrino et al. 2011). Given that discrimination can be the outcome of a variety of factors (which may have not been controlled in an experiment) and that differences in speaking rate of the magnitude reported above are perceptible (Quené 2007), it is fair to ask whether speaking rate differences confounded with rhythm class may have been responsible for discrimination.

This seems plausible, considering the outcome of certain studies. For instance, Ramus et al. (2003) report that in their study, Polish was discriminated from both English and Spanish, although Ramus et al. (1999) had classified it as stress-timed on the basis of results from timing-based metrics. Polish, however, is a fast spoken language; e.g. Malisz (2011) reports an average speaking rate of 6.9 syllables/s. This fast speaking rate was inadvertently exaggerated in the Polish stimuli of Ramus et al. (2003) in which vocalic and consonantal intervals were reduced by approximately 10% in order to shorten the overall duration of the stimuli. Thus, it is possible that

Polish was discriminated from both English and Spanish due to its faster speaking rate and not because it falls onto a rhythm class of its own, as Ramus et al. (2003) proposed.

Further, the use of *flat sasasa* implies that timing characteristics are salient to listeners and processed independently of other prosodic components of the speech signal such as F_0. But these assumptions are contradicted by evidence. First, discrimination is possible between languages that belong to the same class, such as English and Arabic (see Komatsu 2007 and references therein), and even between dialects of the same language (e.g. Nazzi et al. 2000; White, Mattys and Wiget 2012). Second, several unrelated experiments have shown that percepts of duration are affected by other prosodic variables. For instance, Kohler (2008) showed that the perception of prominence relies on a complex interplay between the duration and F_0 of stimuli. Cummins (2009: 25) who examines rhythm as entrainment reports data which make him conclude that “no single simple physical variable can be implicated as the basis for coupling among speakers.” Yu (2010) found that F_0 glides can affect percepts of duration, with vowels with flat F_0 being perceived by listeners as shorter than vowels of the same duration but with rising or falling F_0, a result that largely agrees with those of earlier studies such as Lehiste (1976) and Rosen (1977). The interaction between duration and F_0 is also evident in the results of Arvaniti (2012b) from two experiments in which listeners were asked to rate English, German, Greek, Italian, Korean and Spanish for similarity to non-speech trochees. The ratings depended on the transform used: when the stimuli were low-pass filtered (and thus retained some segmental as well as F_0 information) listeners rated English significantly less trochee-like than all the other languages; when the same stimuli were converted to *flat sasasa* listeners rated English, German and Spanish significantly *more* trochee-like than Greek, Italian and Korean. The fact that responses differed depending on the type of signal transform used indicates that listeners cannot isolate timing information and react to it independently of the other prosodic components in the signal (or they would have responded similarly to the two experiments).

Results like those reviewed above have implications for the interpretation of discrimination experiments and the practice of using impoverished stimuli like *flat sasasa* in rhythm research. First, if the processing of timing information is affected by F_0, as the results of Kohler (2008), Yu (2010) and others suggest, then it is important to ask whether it is legitimate to isolate timing in experiments and extrapolate from impoverished stimuli to the processing of timing in real speech (cf. Hawkins 2003 on the consequences of stripping the speech signal of crucial information). At the same time, the neglected role of speaking rate, which may have been confounded with rhythm class or inadvertently manipulated in previous experiments, may provide an alternative explanation of their results as the outcome of speaking rate differences rather than differences in timing associated with rhythm classes.

Clarifying these issues is imperative as in the past decade or so studies on speech rhythm using timing measures in both production or perception have proliferated and have been used for a host of purposes. These include the examination of cross-linguistic differences in rhythm and the perception of such differences (e.g. Ramus et al. 1999; Grabe and Low 2002; Ramus et al. 2003; White et al. 2012), the rhythm classification of languages (e.g. Cho 2004; Keane 2006), the characterization of atypical speech (e.g. Liss et al. 2009), L2 speech (e.g. White and Mattys 2007; Mok and Dellwo 2008) and child language (e.g. Payne et al. 2011), the study of sociolinguistic and dialectal variation (e.g. Szakay 2008; Giordano and D'Ana 2010; Torgersen and Szakay 2011), and the process of language acquisition itself (Ramus et al. 1999).

The role played in discrimination experiments by speaking rate, F_0 and timing (as exponent of rhythm class) is addressed here by means of five AAX experiments in which stimuli from English were compared to stimuli from Danish, Greek, Korean, Polish and Spanish. The stimuli either retained their original speaking rate and F_0 or were manipulated so that this information was removed from the signal in a way that did not affect timing. It was hypothesized that if discrimination is based on rhythm class, then Polish and Danish which are stress-timed (Ramus et al. 1999 on Polish; Hansen and Pharao 2010 on Danish) would not be discriminated from English, but that Spanish, Greek and Korean would be since they are typically classed as syllable-timed (Dauer 1983 on Spanish and Greek; Ramus et al. 1999 and Grabe and Low 2002 on Spanish; Kim et al. 2008 on Korean). The alternative hypothesis was that discrimination relies on speaking rate and is independent of rhythm class; if so, then fast spoken languages such as Greek and Polish (Dauer 1983; Malisz 2011) should be discriminated from English; in contrast, Korean and Danish, whose speaking rates are very close to English (Thorsen 1980; Kim 2009) should not be. Finally, it was hypothesized that F_0 modulation by providing additional information would aid discrimination independently of the speaking rate manipulation.

2. Methods

Five AAX experiments were run in which context (AA) was always English, and one of the other languages, Danish, Spanish, Greek, Korean or Polish, was the test language (X). Listeners heard two context stimuli followed by one test stimulus and had to decide whether the third stimulus belonged to the same language as the first two.

2.1. *Participants*

The sentences on which the stimuli were based were elicited from one male and one female talker of Danish, Greek, Polish, Spanish and Korean, and two male and two female English talkers. For English, the data of one male and one female talker provided context stimuli and those of the other two provided controls (English stimuli used as tests). The talkers, who were all volunteers, were monolingual native speakers of each language and used the same broadly defined dialect: Southern Californian English, Border Mexican Spanish (Spanish, as spoken in the San Diego-Tijuana border), and the standard varieties of Greek, Polish, Danish and Korean (with the exception of the female speaker of Danish who came from Aarhus but used Standard Danish for the recording). The age of the speakers ranged from 19 to 51 years (mean = 32).

A total of 139 listeners took part in the five experiments. Due to technical difficulties, 18 of them could not complete the task; another 5 participants did not provide responses to more than 4 trials per condition and for that reason their results were removed from further analysis. Results are presented here for 116 participants, 24 per experiment for Greek, Polish and Spanish, and 22 for Korean and Danish. Participants were between 18 and 27 years old (with one outlier at age 43), and were undergraduates at UCSD who took part in the study for course credit. Sixty five of them were monolingual native speakers of English, and 51 were bilinguals. The latter group consisted of simultaneous bilinguals and early sequential bilinguals who were introduced to English before the age of 12 (median age of introduction was 5). There was only one L2 speaker in the group, the age outlier who took part in the Danish experiment; her results were included as they did not deviate from those of the other participants. The linguistically varied background of participants should have little bearing on the results for two reasons: first, both monolingual and bilingual speakers took part in each experiment with no obvious demographic differences across experiments (see Appendix I for a breakdown of participants' linguistic background by experiment); second, results of experiments on rhythm perception have shown no significant differences based on the linguistic background of listeners (e.g. Ramus et al. 2003; Arvaniti 2012b; White et al. 2012).

Neither talkers nor listeners reported any history of speech or hearing problems. Prior to taking part in the recording or experiment, participants signed institutional IRB forms and provided relevant demographic information.

2.2. *Stimuli*

For all languages except Greek and Korean the sentences on which the stimuli were based were recorded in the sound treated booth of the UC San Diego (UCSD) Speech Lab. Talkers read the materials of their language three times in randomized order off a Powerpoint presentation at a speaking rate deemed comfortable for them. The Greek talkers were recorded under similar conditions, but using the facilities and recording studio of the Institute of Speech and Language Processing (ISLP) in Athens. The Korean talkers were recorded in a quiet room in their own home. The number of sentences varied by language and talker (a difference to do mostly with talker availability); e.g. the female Danish talker read a corpus of ten sentences while the male talker read a corpus of thirty sentences.

For the test languages (Danish, Greek, Korean, Polish and Spanish) eight sentences per speaker were selected and used in all experimental conditions for a total of sixteen sentences per test language (see Appendix II). The criteria for selection were that the sentences sound fluent and natural and be of comparable speaking rate across the two speakers of each language. These speaking rates were within the typical ranges reported for each language (e.g. Thorsen 1980 on Danish; Dauer 1983 on English, Spanish, Greek; Kim 2009 for Korean; Malisz 2011 on Polish);

descriptive statistics on speaking rate for the stimuli can be found in Table 1 (for full details see Appendix III; note that reported rates pertain to stimuli not the original sentences from which those were derived; for further details see below).

For English a much larger corpus was collected, since each experiment required five times as many English sentences (used for context and controls) as there were sentences of the test language. In addition, English sentences had to vary more in number of syllables in order to compensate for the fact that speaking rate manipulations (discussed in more detail below) affected the overall duration of stimuli; indicative sentences of English are provided in Appendix II. The English sentences used for stimuli were selected by the same criteria employed for the other languages.

The original sentences were converted to *sasasa* using STRAIGHT in Matlab (http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html; we are grateful to Spyros Raptis, ISLP, and his colleagues for synthesizing the stimuli). Prior to synthesis, silences at the beginning and end of the speech files and any pauses were excised; utterance-initial vowels were also excised so that all stimuli would start with [s]. The diphones used for synthesis came from the stressed syllable of the Greek name ['sasa] recorded utterance-initially in a sentence and elicited from one male and one female speaker of Standard Greek (different from those who recorded the stimuli). Greek diphones were used for synthesis to ensure a lack of bias for the English-speaking listeners: the Greek [s] and [a] are sufficiently different from their English counterparts (Arvaniti 2007) that the stimuli would sound unfamiliar to the participants and thus would make the premise of the experiments, namely that foreign languages were involved, more plausible.

The *sasasa*-converted sentences were further manipulated to create the stimuli for the four conditions of each experiment, *SR-*F_0, *SR-no*F_0*, noSR-*F_0 and no*SR-no*F_0 (where *SR* stands for speaking rate expressed in vocalic intervals per second). The aim of conditions *SR-*F_0 and *SR-no*F_0 was for stimuli to retain their original speaking rate. More precisely, in these conditions the stimuli retained the original durations of their consonantal and vocalic intervals; since vocalic intervals largely correspond to syllable nuclei, it was anticipated that in these conditions the original speaking rates would remain generally unaltered. However, several sentences contained vowels in hiatus and as a result the speaking rates of the stimuli were in some cases different from those of the original sentences (for the speaking rates of the stimuli in the *SR* conditions see Table 1). *SR-F_0* and *SR-noF_0* stimuli differed in terms of F_0; in condition *SR-F_0*, the stimuli retained their original F_0 modulation except that F_0 was adapted in range to that of the appropriate diphone talker so as to minimize perceptual discrepancies between formant frequencies and pitch range. In condition *SR-noF_0*, F_0 was "flattened" to slightly declining F_0 that spanned the middle third of the appropriate diphone talker's range in the utterance from which their diphone was extracted; this range was 97-120 Hz for the male talker and 220-268 Hz for the female talker.

In conditions *noSR-F_0* and *noSR-noF_0*, the same F_0 manipulation was done as before: in *noSR-F_0*, the stimuli retained their original F_0 modulation (adjusted as noted above), while in *noSR-noF_0*, F_0 was "flattened". In addition speaking rate differences between context (English) and test language were eliminated. This was done by multiplying by a factor specific to each stimulus the durations of all its vocalic and consonantal intervals so as to convert the speaking rate of each stimulus to an average derived from the original speaking rate of English and the test language in each experiment. Note that in the two *noSR* conditions the changes in the duration of consonantal and vocalic intervals were proportional; consequently these changes affected only speaking rate, while the *timing characteristics* of the stimuli—that is the variability and relative duration of vocalic and consonantal intervals—remained stable. This means that if rhythm class affiliation as reflected in timing is responsible for discrimination, the elimination of speaking rate differences should not affect discrimination, since the characteristics of the consonantal and vocalic intervals are not affected by the changes in speaking rate (see e.g. Appendix III for %V, a metric argued by Ramus et al. 1999 to predict discrimination, and VarcoC, which in combination with %V has been argued by White and Mattys 2007 to predict rhythm class affiliation in production; here only the role of %V is discussed in some detail since VarcoC showed few differences between languages and has not been connected to a testable hypothesis with respect to rhythm class perception).[2]

One consequence of the elimination of speaking rate differences in the *noSR* conditions was that it was not possible to control the overall duration of stimuli as was done in previous experiments (e.g. Ramus et al. 2003): the shrinking or stretching necessary to achieve the changes in speaking rate meant that the overall duration of the stimuli was unavoidably altered. In order to avoid discrimination due to differences in overall duration between English and test language two steps were taken. First, stimuli for each trial were selected so that they did not differ in overall duration by more than 300 ms, a durational difference expected to be perceptually irrelevant (Lehiste 1977). Second, overall durations varied across trials so that overall stimulus duration could not serve as a discrimination cue. In order to achieve both similarity of overall duration within trials and variability across while maintaining some stability across conditions two further steps were taken: first, the same sixteen stimuli of the test language were used in all conditions; second, the English context and control stimuli were allowed to vary in terms of number of syllables and overall duration, though the English stimuli used within each experiment overlapped substantially across conditions. Post-hoc statistical analysis showed that due to the manipulations involved in the preparation of stimuli and the changes in speaking rate, in some languages and conditions, differences in speaking rate between context and text were smaller than anticipated while differences in sentence duration or syllable count were sizeable (see Table 1); relevant measurements are presented in detail in Appendix III and discussed in section 4 in light on the results.

Each experiment was organized into four blocks of eight trials per condition, four test trials and four controls (i.e. all English). The order of the blocks and conditions was counterbalanced across participants. In each trial the two different context stimuli were selected so that both a female and a male talker were heard; talker gender order was counterbalanced across trials. Each trial was introduced by a 500 Hz tone of 200 ms duration; this was followed by 200 ms of silence, after which the listener would hear the first context sentence while seeing “sentence 1” on screen; after 1 s. of silence the second context sentence was heard while “sentence 2” appeared on screen; the procedure was repeated for the test sentence while “sentence 3” appeared on screen. The test sentence was followed by the on-screen prompt “Press Y if same Press N if different” and listeners had 3 s. to respond before the tone would begin the next trial. Each experiment lasted 15-20 minutes.

Table 1: Means and standard deviations (in parentheses) of various timing measures separately for context and test stimuli in conditions *SRF_0* and *$SR\text{-}noF_0$*; starred values indicate that test stimuli were significantly different [$p < 0.05$] from context stimuli (based on results of ANOVAs with sentence type—context, test—and condition—*SRF_0/$SR\text{-}noF_0$, $noSRF_0$/$noSR\text{-}noF_0$*—as independent variables; one-way ANOVAs with sentence type as the independent variable were used to calculate differences in speaking rate).

Test Language		Contexts (AA)	Test (X)
Danish			
	speaking rate	5.02 (0.45)	5.1 (0.5)
	number of syllables	9.6 (1.6)	9.6 (1.5)
	stimulus duration	1900 (219)	1901 (259)
	%V	42.8 (8.0)	45.7 (4.8)*
Greek			
	speaking rate	5.3 (0.6)	5.9 (0.5)*
	number of syllables	8.1 (1.7)	10.4 (1.5)*
	stimulus duration	1529 (286)	1779 (255)*
	%V	44.6 (7.5)	45.3 (6.1)
Korean			
	speaking rate	4.6 (0.6)	5.0 (0.4)*
	number of syllables	10.4 (1.3)	11.6 (1.3)*
	stimulus duration	2256 (248)	2323 (227)*
	%V	42.2 (6.11)	53.9 (5.8)*
Polish			
	speaking rate	5.2 (0.73)	5.7 (0.6)*
	number of syllables	8.4 (1.6)	11.8 (1.7)*
	stimulus duration	1614 (243)	2084 (381)*
	%V	42.6 (6.5)	40.9 (4.4)
Spanish			
	speaking rate	5.3 (0.5)	5.5 (0.6)
	number of syllables	8.7 (1.7)	11.4 (2.2)*
	stimulus duration	1630 (289)	2049 (245)*
	%V	42.2 (5.8)	46.3 (4.3)*

2.3. *Procedure*

Participants were tested in the sound-treated booth of the UCSD Speech Lab using E-prime on a Windows 7 platform. At the beginning of the experiment the listeners were given instructions that they would hear groups of three sentences in which all consonants had been replaced by [s] and all vowels by [a]. They were also told that the first two sentences of every group came from the same (foreign) language and their task was to decide if the third sentence came from the same language or not. The languages were not named so as not to bias the listeners about the possible characteristics of the languages involved in each experiment (cf. the terms *Sahatu* and *Moltec* of Ramus and Mehler, 1999, and Ramus et al., 2003, which allude to simple vs. complex syllable structure respectively). The participants were informed that their response time was limited and that they were expected to provide an answer even if they were uncertain.

Participants completed a practice session before the experiment. The practice session lasted for a single block of eight trials. Afterwards participants proceeded to the main task (i.e. they were not required to reach criterion). Before the main experiment begun they were encouraged to ask any remaining questions about the task. At the end of the experiment they were given a short questionnaire as an exit interview (questionnaires were not given to the participants in the Polish experiment who had an oral exit interview with the second author instead).

2.4. *Measurements and statistical analysis*

As in previous studies, the results showed conservative bias, i.e. reluctance on the part of the listeners to respond that stimuli were different (see Table 2). Conservative bias made the use of A' scores, the nonparametric analog of *d'*, more suitable for analysis, since conservative bias results in negative *d'* values which are largely uninterpretable (Snodgrass, Levy-Berger, and Haydon 1985). A' scores were calculated for each participant by using hit rate (number of times when the participant correctly answered *different* when the languages were different) and false alarm rate (number of times the participant incorrectly answered *different* when the languages were the same). An A' value of 0.5 indicates chance level while values above 0.5 indicate discrimination. In addition, B''_D which measures bias (Donaldson 1992) was calculated in order to examine bias in more detail. Finally reaction time (RTs) were measured from the end of the last stimulus in each trial; since responses were expected to be conservative (cf. Ramus et al. 2003), RTs could serve as an additional diagnostic regarding the difficulty of conditions and the manner in which participants responded to the task.

The A' values were statistically analyzed in two ways. First, single sample *t*-tests were performed to see whether discrimination was above chance for each language and condition; the same procedure was used for values pooled over languages and values pooled over conditions. Bonferroni correction was applied for tests within each language ($p < 0.0125$), for experimental conditions pooled over languages ($p < 0.0125$) and for languages pooled over conditions ($p < 0.01$). Second, ANOVAs were run with A' scores as the dependent variable, condition as a repeated-measures factor with four levels (*SR-F*$_0$, *SR-noF*$_0$, *noSR-F*$_0$, *noSR-noF*$_0$) and language as a categorical predictor. An ANOVA with the same design was also used to analyze B"$_D$. Finally, ANOVAs were used to analyze RTs using response type (hit, false alarm, miss—in which participants answer *same* when the languages are different—and correct rejection--in which participants answer *same* when the languages are the same) and condition (*SR-F*$_0$, *SR-noF*$_0$, *noSR-F*$_0$, *noSR-noF*$_0$) as repeated measures factors and language as categorical predictor. (Initial ANOVAs on A' and B"$_D$ which included the order in which conditions were presented as a between subjects factor showed no order effect [$F < 1$ in both cases], so order was removed from further analysis.)

Table 2: A' scores for each language and condition and *p*-values for single sample *t*-tests by condition (pooled over languages), by language (pooled over conditions) and separately for each language and condition (for pooled A' scores see Figures 1 and 2).

Language		*SR-F*$_0$	*SR-noF*$_0$	*noSR-F*$_0$	*noSR-noF*$_0$	Pooled over conditions
Danish	A'	0.59	0.55	0.48	0.56	
	p	0.01	n.s.	n.s.	n.s.	0.01
Greek	A'	0.76	0.76	0.58	0.53	
	p	0.001	0.001	0.01	n.s.	0.001
Korean	A'	0.58	0.64	0.54	0.61	
	p	n.s.	0.001	n.s.	0.01	0.001
Polish	A'	0.60	0.56	0.45	0.45	
	p	0.01	n.s.	n.s.	n.s.	n.s.
Spanish	A'	0.50	0.53	0.49	0.54	
	p	n.s.	n.s.	n.s.	n.s.	n.s.
Pooled over languages	*p*	0.001	0.001	n.s.	0.01	

3. Results

3.1. *A' scores*

As can be seen in Table 2, A' scores pooled over conditions were significantly different from chance in the experiments involving Danish, Greek and Korean, a group that includes one stress-timed and two syllable-timed languages. Overall, three experimental conditions were discriminated significantly better than chance across languages, *SR-F_0, SR-noF_0* and *noSR-noF_0*, supporting the idea that speaking rate is important for discrimination (though not the only contributor). When the data are examined separately for each language and experimental condition, at least some conditions in each experiment except that involving Spanish were discriminated better than chance.

The results of the ANOVA largely supported these findings, showing main effects of language [$F(4, 111) = 16.3$, $p < 0.001$] and condition [$F(3, 333) = 16.5$, $p < 0.001$]. Planned comparisons showed that Greek was overall better discriminated than any of the other languages [$p < 0.0001$ for comparisons with Danish, Polish and Spanish; $p < 0.01$ for the comparison with Korean]. Korean had the second highest pooled A' value which was significantly higher than the scores of Danish, Polish and Spanish [$p < 0.05$ for Danish; $p < 0.001$ for Polish and Spanish]. Danish, Polish and Spanish had the lowest scores which were not significantly different from each other (see Figure 1). Planned comparisons also showed that the two conditions retaining original speaking rate (*SR-F_0* and *SR-noF_0*) were similar to each other and significantly better discriminated than the *noSR* conditions [$p < 0.0001$ for all pairwise comparisons] which were also comparable to each other (see Figure 2).

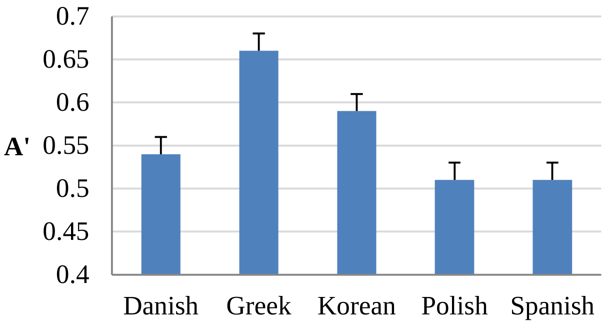

Figure 1: Mean A' and standard errors by language.

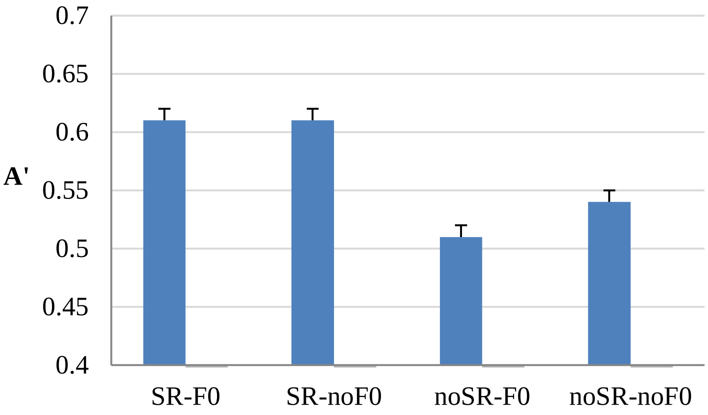

Figure 2: Mean A' and standard errors by experimental condition.

3.2. *Bias*

The ANOVA on B"$_D$ values showed effects of language [$F(4,111) = 3.9$, $p < 0.01$] and condition [$F(3, 333) = 66.7$, $p < 0.001$]. Planned comparisons indicated that the language effect was primarily due to Danish which showed significantly more conservative bias than Greek [$p < 0.001$], Polish [$p < 0.01$] and Korean [$p < 0.05$] as illustrated in Figure 3; Spanish also showed relatively conservative bias though only the comparison with Greek was close to significance [$p = 0.057$]. As shown in Figure 4, the effect of condition was such that bias increased in the order *SR-F$_0$* < *noSR-F$_0$* < *SR- noF$_0$* < *noSR-noF$_0$* [$p < 0.0001$ for all comparisons, except *SR-F$_0$* vs. *noSR-F$_0$*, where $p < 0.05$, and *SR-F$_0$* vs. *noSR-noF$_0$* where $p < 0.001$]. In sum, there was a strong conservative bias in the conditions that did not include *F$_0$* information as compared to those that did, while Danish, the language closest to English in terms of speaking rate showed the most conservative bias. Crucially, condition *SR-F_0*, the condition that included both speaking rate and F_0 information showed the most liberal bias of all, suggesting that in this condition listeners felt more confident in their judgments and in hearing differences among stimuli.

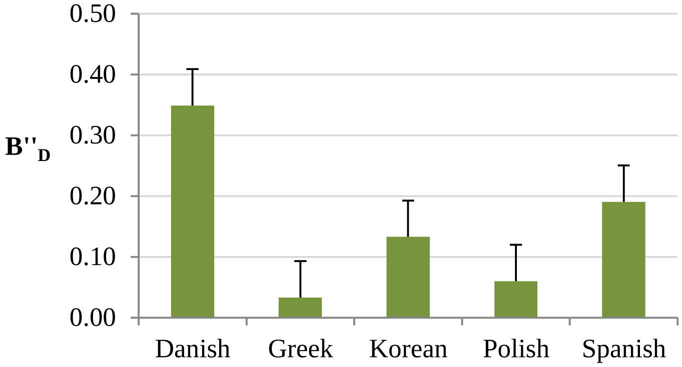

Figure 3: B"$_D$ values and standard errors by language.

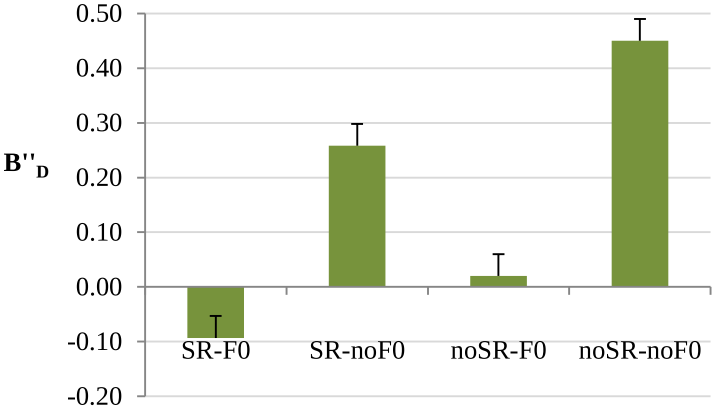

Figure 4: B"$_D$ values and standard errors by experimental condition.

3.3. *Reaction times*

RTs showed effects of both response type [$F(3,333) = 6.04$, $p < 0.001$] and condition [$F(3,333) = 6.03$, $p < 0.001$] that supported the discrimination results. For response, planned comparisons showed that hits were responded to similarly to correct rejections and faster than misses and false alarms [$p < 0.001$]; correct rejections were also responded to faster than misses [$p < 0.05$]; see Figure 5. Thus, overall, accurate responses were reached faster than inaccurate ones, a difference that lends credence to the discrimination results. As illustrated in Figure 6, planned comparisons showed that *noSR-F_0* had significantly longer RTs than all other conditions [$p < 0.05$], a result that agrees with the low discrimination scores for this condition which was clearly the most difficult for participants.

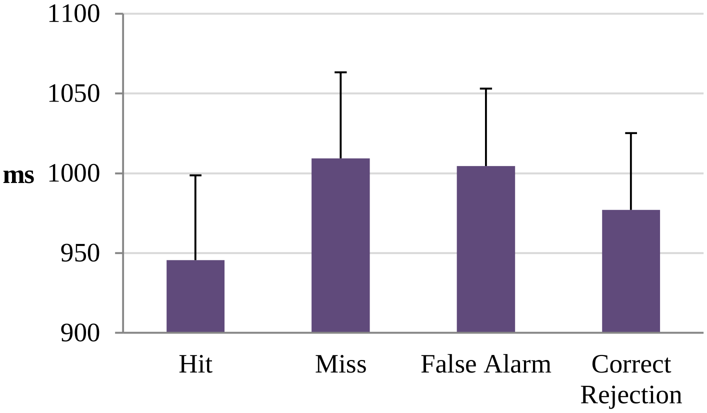

Figure 5: RTs and standard errors by type of response.

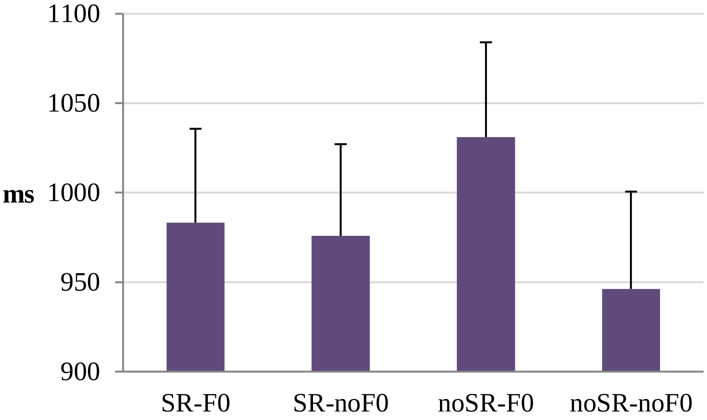

Figure 6: RTs and standard errors by experimental condition.

4. Discussion

The series of experiments presented here aimed at testing whether it is differences in rhythm class or other aspects of prosody, specifically speaking rate and F_0, that drive discrimination in experiments using the *sasasa* transform. The results provide support for the latter hypothesis but also reveal that the answer is far from simple.

First, the data showed that discrimination is not based on differences between rhythm classes; if that had been the case, then neither Danish nor Polish should have been discriminated from English, while Spanish should have had the highest discrimination scores. On the contrary, both Danish and Polish were discriminated from English at least when both speaking rate and F_0 information were present, while Spanish was not discriminated in any condition. In addition, Korean which is also classed as syllable-timed should have been consistently discriminated from English but was not.

Unlike rhythm class, speaking rate did play a consistent role in discrimination. As noted, experimental conditions with original speaking rate showed overall better discrimination, faster reaction times and less conservative bias than conditions in which speaking rate differences were eliminated, overall suggesting that these were the conditions in which cross-linguistic differences were most easily and reliably perceived. This was most strongly demonstrated for Greek which showed the highest discrimination and had the biggest speaking rate difference as compared to English. Spanish, on the other hand, was not discriminated from English, a result that can plausibly be linked to the fact that its stimuli were of a speaking rate comparable to the English context stimuli.

F_0 also played a part, though its role was not entirely consistent across languages and experimental conditions. In Polish it aided discrimination *in combination with* speaking rate (which was significantly higher in Polish than English), while in Danish F_0 is the most plausible cause behind discrimination in the *SR-*F_0 condition, since differences in speaking rate between Danish and English were negligible. Discrimination on the basis of F_0 alone is evidenced in Greek as well in the *noSR-*F_0 condition. However, F_0 did not always enhance discrimination as originally hypothesized; rather the results suggest a complex interaction with durational characteristics including speaking rate. This conclusion is supported by the fact that across experiments the *noSR-*F_0 condition had low A' values and the longest RTs, indicating that the lack of speaking rate differences together with the effect that F_0 modulation had on perceived durations confused rather than helped listeners. This was most prominent in Korean which was discriminated from English only in the two conditions that did not include F_0 modulation (*SR-noF*$_0$ *and noSR-noF*$_0$). For *SR-noF*$_0$ at least, a plausible explanation is that the reliable but small differences in speaking rate between Korean stimuli and English were not sufficiently salient in the presence of F_0 modulation but came to the fore when F_0 modulation was eliminated.

The fact that discrimination in Korean was also possible in *noSR-noF$_0$* indicates that timing differences beyond speaking rate also played a part in discrimination. Korean, as shown in Table 1, had a significantly higher %V (percentage of speech made up of vocalic intervals) than English; this was most likely the outcome of the extensive final-lengthening that affected primarily utterance-final vowels in Korean. These long final vowels provided information to listeners that allowed them to discriminate Korean from English in the absence of other prosody components. Clearly, however, this feature was not as easily detectable when speaking rate and F_0 information were present or Korean would have been discriminated in all conditions.

Since %V is a measure of timing that is said to predict discrimination (Ramus et al. 1999; Ramus et al. 2003), it is worth considering whether differences in %V were behind discrimination in the present experiments. This possibility seems unlikely for two reasons. First, the perceptual significance of %V is unclear as no experiments to our knowledge have reported just noticeable differences in %V (or any other rhythm metric). A liberal approach to this matter would be to assume that any statistically significant differences are perceptible and thus lead to discrimination, but the present results do not support this possibility in that %V differences were not consistentently related to discrimination. Danish stimuli had significantly higher %V than English, yet discrimination was not possible in most conditions; similarly, %V differences did not lead to discrimination in the Spanish experiment or in the conditions with F_0 modulation in the Korean experiment. In contrast, comparable %V values between Greek and English did not hinder discrimination; the same applies (to an extent) to Polish as well. This inconsistency in the role of %V suggests that most likely it was localized final lengthening that led to discrimination in Korean *noSR-noF$_0$*, and that this localized difference was reflected in %V a global timing measure (cf. Arvaniti 2009 on the effects of specific timing features on timing-based measures, and White et al. 2012 on similar conclusions about the role of localized timing differences in discrimination).

Similarly unlikely is the role of differences in syllable count and overall stimulus duration observed in some of the present experiments. Although some of these differences between contexts and stimuli were statistically significant, it is doubtful they were of a magnitude that would render them perceptually relevant; e.g. overall stimulus duration differences in Greek and Korean were less than 15%, the typical JND for duration (Lehiste 1977). The perceptual significance of syllable count, on the other hand, is unclear. Crucially, however, the relation between these durational characteristics of the stimuli and discrimination was as inconsistent as that of %V. Thus, even though one could attribute discrimination in Greek to the longer duration of its sentences and the higher syllable count compared to English, larger (and clearly perceptually relevant) differences in these two variables did not lead to discrimination in the Polish and Spanish experiments making it difficult to sustain the argument that it was these differences that caused discrimination (see Table 1 and Appendix III for details).

Overall then the results point to the fact that any differences in timing were probably local ones and insufficient *on their own* to consistently lead to discrimination, while their role was minimized when speaking rate and F_0 information was present. Like other results, this suggests that timing was not processed independently of the other prosodic features present in the experiments discussed here; if it were we would not see results being swayed by the presence of speaking rate or F_0 differences (see also Arvaniti, 2012b, for similar conclusions).

The above does not mean to preclude the possibility that all the variables discussed (speaking rate, F_0, syllable count, stimulus duration, %V) played some part in one condition or another in some experiments, given the overall difficulty of the task. As noted, A' values were generally not particularly high (Figures 1 and 2), while conservative bias was strong (Table 2 and Figures 3 and 4), suggesting that participants could not reliably hear differences between languages. Difficulty is also reflected in participant ratings in the exit questionnaires: on a difficulty scale from 1 to 7 the mode across experiments was 6. Similar results have been reported before; e.g. Ramus et al. (2003) used the same paradigm and report conservative bias and A' values ranging from 0.48 (Catalan vs. Spanish) to 0.74 (Polish vs. Spanish).

The difficulties that listeners encountered should not be overlooked when discrimination results are considered. First, the difficulty of the task makes it likely that participants utilize whatever information is available to them, however inadvertently, and that therefore the interpretation of results from discrimination experiments is far from easy to attribute exclusively and reliably to one variable. Second, the difficulties suggest that the task itself—seen by some as the foundation of language acquisition (e.g. Ramus et al. 1999)—is not as easy and natural as it has been suggested.

Additional repercussions relate to the use of *sasasa*, a transform that removes prosodic information, and the primacy of timing in perceiving rhythm class distinctions. First, given that speaking rate, F_0 and timing cannot be absent from real speech, the present results suggest that the *sasasa* transform is not ecologically valid since the information it preserves is typically influenced in perception by information it removes, such as F_0 (Lehiste 1976; Rosen 1977; Yu 2010). If so, then the present experiments cast doubt on the claim that the relative timing of consonantal and vocalic intervals helps listeners discriminate between languages of different rhythm classes since supporting results have only been obtained with stimuli stripped of other information (e.g. Ramus & Mehler 1999; Ramus et al. 1999; Ramus et al. 2003). The present results further suggest that any timing differences, in terms of interval durations, are probably closely linked to differences in speaking rate as well.[3] This in turn implies that, at a minimum, discrimination with *flat sasasa* should be treated with caution, particularly when used to deterimne the rhythm classification of a languages (cf. Cho 2004 who concluded on the basis of a series of such experiments that Korean is mora-timed, a classification that is at odds both with the phonological structure of the language and all other studies on its rhythm so far).

In short, results from both production and perception have not provided support for the view that timing is the cornerstone of speech rhythm (Barry, Andreeva, and Koreman 2009; Loukina et al. 2011; Arvaniti 2012a; Arvaniti 2012b). At the same time, existing results that appear to support this view are amenable to alternative and at least equal plausible interpretations; e.g. discrimination results are likely due to differences in speaking rate that were inadvertently not controlled in previous experiments. Given the above, the linguistic view of speech rhythm as resting on timing relations and reflecting a typology of three timing-based classes seems hard to sustain. Since discrimination experiments have been used to bolster the rhythm class hypotheses, the present results which found no evidence that rhythm class was pivotal in discrimination cast further doubt on the idea of rhythm classes as a valid typological distinction.

4.1 Towards an alternative view of speech rhythm

If rhythm classes do not provide an adequate understanding of speech rhythm, it is worth briefly considering what could form an alternative basis for rhythm research. Arvaniti (2009: 57) suggested that a definition based on the psychological understanding of rhythm as "the perception of series of stimuli as series of groups of similar and repetitive pattern" be adopted in linguistic research (Woodrow 1951; Fraisse, 1963; 1982). Adopting this definition would mean investigating the possibility that rhythm is cross-linguistically based on the grouping of prosodic constituents (such as syllables) created by patterns based on the relative prominence of these constituents. This view is compatible with the definition of rhythm provided above and opens up the possibility that rhythm may be based on a variety of prosodic components that help create differences in relative prominence among relevant constituents.

Recent work by Arvaniti and colleagues (Arvaniti and Tilsen 2011; Chung, 2011; Chung and Arvaniti 2012; Tilsen 2011; Tilsen and Arvaniti subm.) provides initial evidence to this effect and is thus presented here in brief. In one strand of this research, Tilsen and Arvaniti used Empirical Mode Decomposition on the corpus of Arvaniti (2012a) to uncover periodicities in the amplitude envelope of the speech signal. Their research shows that the instantaneous frequencies of the first two modes have similar means in all the languages in the corpus (English, German, Greek, Italian, Korean and Spanish): 5.7–6.7 Hz for the instantaneous frequency of the first mode (ω_1) and 2.3–2.6 Hz for the second mode (ω_2). These frequencies are within the Theta and Delta bands (4–10 Hz and 1.5–4 Hz respectively) of oscillating networks of neurons which have been shown to be crucial for speech processing and rhythm entrainment (among many, Luo and Poeppel 2007; Goswami 2011; see also Goswami & Leong, this volume). From a linguistic perspective, the frequencies reported by Tilsen and Arvaniti correspond well to the periodicities of vowel nuclei and feet respectively, suggesting more similarities than differences across the six languages in their corpus, despite the different prosodic structures and assumed rhythm class affiliations of these languages. For example, these general patterns apply both to English, the quintessential stress-timed language, and Korean, a language that does not have stress or foot structure (Jun 2005). Supporting evidence for Korean comes from the work of Chung and Arvaniti who used speech cycling—in which speakers talk in time with a metronome—to determine whether some syllables are privileged in Korean in a similar fashion to English stressed syllables (cf. Cummins and Port 1998). They found that this indeed applies to the initial syllables of Accentual Phrases, a prosodic constituent that shares similarities both with the foot and the phonological word of other languages (Jun 2005). Their durational data regarding the periodicity of Accentual Phrase-initial syllables are very close to the second mode's instantaneous frequencies reported for Korean by Tilsen and Arvaniti (subm.) and agree well with durational data from Kim (2009), leading credence to the idea that even in Korean, a language without stress and feet, syllables are rhythmically grouped with one syllable in each group becoming salient by virtue of its position. In short, the studies briefly reviewed here provide prima facie evidence that basic periodicities creating prominence-based groupings are present in speech and likely relevant for perception given their frequency ranges.

5. Conclusions

The present series of experiments show that it is not possible to attribute discrimination between languages to some general typological differences in rhythm class encoded as variation in durational intervals or *timing*. Further these experiments show that discrimination with impoverished stimuli is difficult and that listeners take advantage of any differences they can find that will help them in their task. Such differences relate primarily to speaking rate and F_0 but also to localized timing differences, such as final lengthening, when speaking rate and F_0 are absent. Since all these prosodic factors are ordinarily present in the speech signal and interact in perception, the present results cast doubt on the view that timing is primary and can be processed by listeners independently all the other prosodic variables. As such, the results have repercussions both for experimental practices and our understanding of rhythm. First, they show that eliminating prosodic information from stimuli results in ecologically invalid manipulations hence dubious results. Second, they add to the body of research showing that neither production nor perception supports the idea of speech rhythm as a typology encoded in timing and by extension to the concept of timing-based rhythm classes overall.

Acknowledgements

We thank the members of the Institute for Language and Speech Processing (ISLP), Athens, Greece, especially Spyros Raptis and his colleagues who prepared the *sasasa* stimuli and arranged for the Greek recordings to take place in the ISLP studio. Thanks are also due to our talkers and to our language consultants Nicholas Frickelton (Danish), Sarah Choi (Korean), Bożena Pająk (Polish) and Cabriella Caballero (Spanish). The financial support of the UCSD Committee on Research through Grant no. LIN054G with Amalia Arvaniti as PI and Tara Rodriquez as R.A. is hereby gratefully acknowledged.

Notes

1. ΔC refers to the standard deviation of consonantal intervals in a given stretch of speech, while %V refers to the vocalic percentage (Ramus et al. 1999). VarcoC and VarcoV are normalized standard deviations of consonantal and vocalic intervals respectively, i.e. standard deviations divided by the mean (Dellwo 2006). PVIs or Pairwise Variability Indices are means of absolute differences between successive vocalic or consonantal intervals presented either *raw* (rPVI) or in normalized form (nPVI) by dividing the difference of each pair by their mean (Grabe and Low 2002).

2. A recent practice dictates that one provides a mass of timing-based metrics selecting suitable results for discussion (e.g. White et al. 2012). As amply demonstrated in Arvaniti (2009, 2012a) it is unclear what different metrics reflect, how they relate to each other and what their connection is to perception, if any. Here, as noted earlier, only %V is discussed in relation to the results of the experiments, since it is one of the few measures that is relatively transparent and has been explicitly argued to predict discrimination (Ramus et al. 1999).

3. An alternative possibility is that the task here is unrepresentative of discrimination between languages, since listeners had to reach decisions based on three utterances (at a time) that may

not have been representative of each language's rhythm as a whole. If so, this possibility puts into question all discrimination experiments on speech rhythm, since they are all based on the same paradigm. A thorough discussion of this possibility is beyond the scope of this paper; rather the present results are interpreted in light on prevailing assumptions within the relevant literature. For example, an implicit assumption of most timing-based studies arguing for the rhythm class hypothesis—much as this assumption defies empirical evidence (Arvaniti 2012a)—is that there is little variation within each language and thus even small samples are representative of the language itself; thus many production and perception studies have relied on small samples either in terms of number of speakers, or in terms of numbers of stimuli, or both (e.g. Ramus et al. 1999 relied on four speakers per language and five sentences per speaker, Grabe & Low 2002 recorded one speaker per language, while the discrimination experiments of White et al. 2012 are based on five English and five Spanish stimuli).

References

Abercrobie, David. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Arvaniti, Amalia. 2007. Greek phonetics: the state of the art. *Journal of Greek Linguistics* 8. 97–208.

Arvaniti, Amalia. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66. 46–63.

Arvaniti, Amalia. 2012a. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40. 351–373.

Arvaniti, Amalia. 2012b. Rhythm classes and speech perception. In O. Niebuhr (ed.), *Understanding Prosody: The Role of Context, Function and Communication*, 75-92. Berlin/Boston: De Gruyter.

Arvaniti, Amalia and Sam Tilsen. 2011. Comparing envelope- and interval-based rhythm metrics. *Journal of the Acoustical Society of America* 130. 2470.

Barry, William, Bistra Andreeva, and Jacques Koreman. 2009. Do rhythm measures reflect perceived rhythm? *Phonetica* 66. 1–17.

Bertinetto, Pier Marco. 1989. Reflections on the dichotomy <<stress>> vs. <<syllable timing>>. *Revue de Phonétique Appliquée 91-92-93*. 99–129.

Cho, Moon-Hwan. 2004. Rhythm typology of Korean speech. *Cognitive Processing* 5. 249–253.

Chung, Younah. 2011. Speech cycling in Korean. *Journal of the Acoustical Society of America* 130. 2470.

Chung, Younah, and Amalia Arvaniti. 2012. Speech cycling in Korean. Paper presented at Perspectives on Rhythm and Timing (PoRT), University of Glasgow, July 19-21, 2012.

Clopper, Cynthia G. and Rajka Smiljanic. 2011. Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics* 39. 237–245.

Cummins, Fred. 2009. Rhythm as affordance for the entrainment of movement. *Phonetica* 66. 15–28.

Cummins, Fred and Robert F. Port. 1998. Rhythmic constraints on stress-timing in English. *Journal of Phonetics* 31. 139–148.

Dauer, Rebecca M. 1983. Stress-timing and syllable-timing reanalysed. *Journal of Phonetics* 11. 51–62.

Dellwo, Volker 2006. Rhythm and speech rate: A variation coefficient for deltaC. In Pawel Karnowski, and Imre Szigeti (eds.), *Language and Language-Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*, 231–241. Frankfurt am Main: Peter Lang.

Donaldson, Wayne. 1992. Measuring recognition memory. *Journal of Experimental Psychology: General* 121. 275–277.

Fletcher, Janet. 2010. The prosody of speech: Timing and rhythm. In William J. Hardcastle, John Laver, & Fiona E. Gibbon (eds.), *The Handbook of Phonetic Sciences*, 521–602. Oxford: Wiley-Blackwell.

Fraisse, Paul. 1963. *The Psychology of Time*. New York: Harper and Row.

Fraisse, Paul. 1982. Rhythm and tempo. In Diana Deutsch (ed.), *The Psychology of Music*, 149–180. New York: Academic Press.

Giordano, Rosa, and Leandro D'Ana. 2010. A comparison of rhythm metrics in different speaking styles and in fifteen regional varieties of Italian. *Speech Prosody* 2010.

Goswami, Usha. 2011. A temporal sampling framework for developmenta dyslexia. *Trends in Cognitive Sciences* 15. 3–10.

Goswami, Usha, and Victoria Leong. This volume. Speech rhythm and temporal structure: Converging Perspectives? *Laboratory Phonology*.

Grabe, Esther, and Ee Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. In Carlos Gussenhoven, and Natasha Warner (eds.), *Papers in Laboratory Phonology* 7, 515–546. Cambridge: Cambridge University Press.

Hansen, Gert F., and Nicolai Pharao. 2010. Prosody in the Copenhagen multiethnolect. In Pia Quist, and Bente A. Svendsen (eds.), *Multilingual Urban Scandinavia*, 79–95. Bristol: Multilingual Matters.

Hawkins, Sarah. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31. 373–405.

Horton, Russell, and Amalia Arvaniti. (subm.). Language clustering and classification using timing-based rhythm metrics.

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

Keane, Elinor. 2006. Rhythmic characteristics of colloquial and formal Tamil. *Language and Speech* 49. 299–332.

Kim, Jeesun, Chris Davis, and Anne Cutler. 2008. Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and Speech* 51. 343–359.

Kim, Seoncheol. 2009. A preliminary study on the relationship between speech rate and prosodic unit generation in Korean read speech. *Journal of the Linguistic Society of Korea* 53. 225–253.

Kohler, Klaus J. 2008. The perception of prominence patterns. *Phonetica* 65. 257–269.

Kohler, Klaus J. 2009. Rhythm in speech and language: A new research paradigm. *Phonetica* 66. 29–45.

Komatsu, Masahiko. 2007. Reviewing human language identification. *Lecture Notes in Computer Science* 4441/2007. 206–228

Lehiste, Ilse. 1976. Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics* 4. 113–117.

Lehiste, Ilse. 1977. Isochrony reconsidered. *Journal of Phonetics* 5. 253–263.

Liss, Julie M., Laurence White, Sven L. Mattys, Kaitlin Lansford, Andrew J. Lotto, Stephanie M. Spitzer, and John N. Caviness. 2009. Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research* 52. 1334–1352.

Loukina, Anastassia, Greg Kochanski, Burton Rosner, Elinor Keane, and Chilin Shih. 2001. Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America* 129. 3258–3270.

Luo, Huan, and David Poeppel. 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54. 1001–1010.

Malisz, Zofia. 2011. Speaking rate differentiated analyses of timing in Polish. In *Proceedings of 17th International Congress of Phonetic Sciences*, Hong Kong, 17–21 August 2011. 1322–1325.

Mok, Peggy, and Volker Dellwo. 2008. Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Proceedings of the Speech Prosody 2008,* 423–426. Campinas, Brazil.

Nazzi, Thierry, Josiane Bertoncini, and Jacques Mehler. 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24. 756–766.

Nazzi, Thierry, Peter W. Jusczyk, and Elizabeth K. Johnson. 2000. Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language* 43. 1–19.

Nazzi, Thierry, and Franck Ramus. 2003. Perception and acquisition of linguistic rhythm by infants. *Speech Communication* 41. 233–243.

Payne, Elinor, Brechtje Post, Lluisa Astruc, Pilar Prieto, Maria del Mar Vanrell. 2011. Measuring child rhythm. *Language and Speech* 54. 1–27.

Pellegrino, François, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language* 87. 539–558.

Port, Robert F., Jonathan Dalby, and Michael O’Dell. 1987. Evidence for mora-timing in Japanese. *Journal of the Acoustical Society of America* 81. 1574–1585.

Quené, Hugo. 2007. On the just noticeable difference for speaking rate in speech. *Journal of Phonetics* 35. 353–362.

Ramus, François, and Jacques Mehler. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America* 105. 512–521.

Ramus, François, Emmanuel Dupoux, and Jacques Mehler. 2003. The psychological reality of rhythm class: Perceptual studies. *Proceedings of the 15th International Congress of Phonetic Sciences*, 337–342. Barcelona.

Ramus, François, Marina Nespor, and Jacques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73. 265–292.

Renwick, Margaret E. L. 2011. Quantifying rhythm: Interspeaker variation in %V. *Journal of the Acoustical Society of America* 130. 2567.

Rosen, S. M. 1977. The effect of fundamental frequency patterns on perceived duration. KTH Dept. of Speech, Music and Hearing-Quarterly Progress and Status Report 18. 17–30.

Szakay, Anita. 2008. *Ethnic dialect identification in New Zealand: The role of prosodic cues*. Saarbrücken: VDM Verlag.

Snodgrass, Joan G., Gail Levy Berger, and Martin Haydon. 1985. *Human Experimental Psychology*. New York: Oxford University Press.

Thorsen, Nina. 1980. Neutral stress, emphatic stress, and sentence intonation in Advanced Standard Copenhagen Danish. *Annual Report of the Institute of Phonetics, University of Copenhagen* 14. 121−205.

Tilsen, Sam. 2011. Speech rhythm analysis with empirical mode decomposition. *Journal of the Acoustical Society of America* 130. 2470.

Tilsen, Sam, and Amalia Arvaniti. Submitted. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages.

Torgersen, Eivind, and Anita Szakay. 2011. A study of rhythm in London; Is syllable-timing a feature of multicultural London English? *University of Pennsylvania Working Papers in Linguistics 17(2): Selected Papers from NWAVE 39*, 165–174.

Yu, Alan C. L. 2010. Tonal effects on perceived vowel duration. In Cécile Fougeron, Barbara Kühnert, Mariapaola D' Imperio, and Nathalie Vallée (eds.), *Laboratory Phonology 10*, 151–168. De Gruyter Mouton.

White, Laurence, and Sven L. Mattys. 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics 35*. 501–522.

White, Laurence, Sven L. Mattys, and Lukas Wiget. 2012. Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language* 66. 665–679.

Woodrow, Herbert. 1951. Time perception. In Stanley Smith Stevens (ed.), *Handbook of Experimental Psychology*, 1224–1236. New York: Wiley.

Appendix I: Demographic characteristics and language background for the participants in each experiment.

			Participant demographics				
Experiment	L1		Gender			Age statistics	
			F	M	TOTAL		
Danish	Cantonese		1			Range	18–43
	English		7	2		Median	21
	Hebrew		1			Mean	21.3
	Hungarian		1			SD	5
	Korean		4				
	Mandarin			2			
	Spanish		2				
	Vietnamese		2				
		TOTAL	18	4	22		
Greek	Armenian		1			Range	18–24
	Burmese			1		Median	20
	English		13	2		Mean	20.3
	Korean		2			SD	1.5
	Mandarin		2	1			
	Spanish		1				
	Tagalog		1				
		TOTAL	20	4	24		
Korean	Bulgarian		1			Range	18–27
	Cantonese		1			Median	20
	English		9	2		Mean	20.3
	Hebrew		2			SD	2.1
	Japanese			1			
	Mandarin		1				
	Spanish		1				
	Vietnamese		3	1			
		TOTAL	18	4	22		
Polish	Czech			1		Range	18–23
	English		16			Median	20
	Hindi		1			Mean	20
	Korean		3			SD	1.3
	Spanish		2				
	Vietnamese		1				
		TOTAL	23	1	24		
Spanish	English		11	2		Range	19–26
	Korean		2			Median	21
	Mandarin		3			Mean	20.6

Polish	1			SD	1.5
Spanish	4				
Vietnamese	1				
TOTAL	22	2	24		

Appendix II: The sentences on which the *sasasa* stimuli were based; the list is exhaustive for all languages except English for which the appendix includes a representative sample of sentences.

Danish

1. Mads Mikkelsen og Viggo Mortensen er danskere.
2. En gammel mand gik en tur med hans lille hund.
3. En lille pige fik en længe ønsket gave.
4. Is og frugtsalat smager godt på pandekager.
5. Små Lego-klodser er lettere end store klodser.
6. Hovedstaden har været København siden det 15. århundrede.
7. Langs Vestkystens strande er det smukt og dejligt.
8. I forgårs udnævnte statsministeren en ny minister.
9. Luftfartsselskabene lider under vulkanen.
10. En skjorte har mange forskellige farver og mønstre.
11. Mormors julemad er det bedste på hele jorden.
12. Ostemadder med gamle Ole lugter som bare fanden.
13. Om efteråret mister trærne deres gyldne blade.
14. Nytårsaften kræver masser af fyrværkeri.
15. Alle unge mennesker har mobiltelefoner nutildags.
16. Lamper spilder energy hvis tændt om dagen.

English

1. A tomato recall would hurt farmers.
2. Larry found linguistics boring.
3. Our flight was delayed in New York.
4. He married her for the wrong reasons.
5. Manny got some cool birthday presents.
6. Using the elevator during a fire is unsafe.

7. Larry proctored the exam all by himself.
8. San Diego is often colder than people think.
9. Sam cares more for her cat than her friends.
10. Sasha came from Russia two years ago.
11. The San Diego water restrictions are tough this year.
12. I prefer gelato to American ice cream.
13. Skateboarding without a helmet is dangerous.
14. Stefanie will come straight to the restaurant to meet us.
15. Most parents are happy to talk about their children.
16. We’re flying to Cabo first thing in the morning.

Greek

1. Θέλω να δω τις ειδήσεις των εννιά.
2. Ο ήλιος θα καταστρέψει την επιδερμίδα σου.
3. Οι μαρκαδόροι μας έχουν ξεθωριάσει.
4. Η Μαρίνα ταξιδεύει στην Πολυνησία.
5. Η συνεχής συννεφιά με καταθλίβει.
6. Η θάλασσα σήμερα είναι λάδι.
7. Η πολλή βροχή θα καταστρέψει τα σπαρτά.
8. Μ’αρέσει πολύ να ζωγραφίζω.
9. Ο Αργοσαρωνικός είναι για τους τουρίστες.
10. Ο Ευβοϊκός μολύνθηκε από τις περσινές φωτιές.
11. Ο υπολογιστής μου έχει χαλάσει.
12. Μόλις άκουσα το ξυπνητήρι.
13. Άφησα το μολύβι σου δίπλα στο τραπέζι.
14. Αυτή η σακούλα είναι σκισμένη.
15. Η Σάσα αγόρασε καινούργιο πλυντήριο.
16. Θα ήθελα να επισκεφτώ τη Βενετία.

Korean

1. 나는 선물을 받는 것을 좋아해요.

2. 하늘에 아름다운 무지개가 떴어요.
3. 은애는 중요한 약속을 만들었어요.
4. 학생들이 도서관에서 공부를 하네요.
5. 호랑나비가 하늘을 향해 날아갔어요.
6. 여름에는 시원한 수박이 맛있어요.
7. 용수가 부모님께 화를내고나갔나봐.
8. 윤미와 효리가 아침을먹고 있다.
9. 이순신 장군의 거북선은거대하다.
10. 오른쪽에 있는 출구로 나가십시오.
11. 노을이 지는 섬마을이 아름답구나.
12. 은수와 바다로 향해 달리고 싶다.
13. 먹구름이 비 와 바람을 몰고 오는구나.
14. 노란색 꽃이 우리집 마당을 덮었어.
15. 여우가 호랑이한테 시집 가나봐.
16. 연수는 미래를 향해 달리고 있다.

Polish

1. Wieczorem idę do kina na romantyczny film.
2. Siedem razy trzy jest dwadzieścia jeden.
3. Jutro mam bardzo ważny egzamin.
4. Chłopak mojej siostry ma na imię Kuba.
5. Zielony kolor jest bardzo uspokajający.
6. Wybory prezydenckie będą w przyszłym roku.
7. W sobotę idziemy na dyskotekę.
8. Muszę kupić sobie nowy szalik na zimę.
9. Adam jest bardzo dobrym kolegą.
10. Nie wiem co dzisiaj ugotować na kolację.
11. Jutro będzie strasznie gorąco w całym kraju.
12. Nie bardzo mi się podoba ten film.

13. Ręczniki leżą w szafce na dolnej półce.
14. Zastanawiam się, czy wyjazd w góry to był dobry pomysł.
15. Rzadko wychodzimy wieczorami z domu.
16. Mój samochód stoi na parkingu.

Spanish

1. Las telenovelas me gustan mucho.
2. Esta computadora está muy vieja.
3. Su regalo era una bufanda amarilla.
4. Las telenovelas mexicanas me gustan mucho.
5. El Tri obtuvo triumfo histórico ante Francia.
6. Mi mejora amiga es Valentina.
7. Mi hija nació el 2 de febrero.
8. Mi debilidad son los chocolates.
9. Juan busca a Darío para decirle la verdad.
10. La casa estaba cerca de la plaza.
11. Este verano iremos a Puerto Vallarta.
12. El clima de Los Cabos es caluroso.
13. Los meteorólogos pronostican lluvias en el Sur.
14. La exposición de joyas prehispánicas me gusto mucho.
15. Su comportamiento fue muy grosero.
16. Alba trabajaba muchísimas horas cada día.

Appendix III: Mean and standard deviations (in parentheses) of number of syllables, overall stimulus duration, speaking rate and %V for English context sentences (A1 and A2) and test sentences for each language; test values are starred if they were significantly different [$p < 0.05$] from context sentences in a particular condition (based on ANOVAs with sentence type—context, test—and condition—$SRF_0/SR\text{-}noF_0$, $noSRF_0/noSR\text{-}noF_0$—as independent variables; one-way ANOVAs with sentence type as the independent variable were used for differences in speaking rate).

Danish						
Condition	Stimulus	No of sylls	Duration (ms)	Speaking rate	%V	VarcoC
$SR\text{-}F_0$	A1	9.6 (1.3)	1924 (222)	5.0 (0.4)	44.0 (7.8)	35 (2.7)
	A2	9.6 (1.8)	1876 (222)	5.1 (0.5)	41.7 (8.4)	35.7 (2.5)
	Test	9.6 (1.5)	1901 (259)	5.1 (0.5)	45.7 (4.8)*	34.7 (2.8)
$SR\text{-}noF_0$	A1	9.9 (1.7)	1927 (255)	5.1 (0.5)	43.5 (8.4)	3.2 (2.7)
	A2	9.3 (1.3)	1873 (182)	4.9 (0.4)	42.2 (8.0)	35.5 (2.6)
	Test	9.6 (1.5)	1901 (259)	5.1 (0.5)	45.7 (4.8)*	34.7 (2.8)
$noSR\text{-}F_0$	A1	9.6 (1.5)	1931 (295)	4.95	42.8 (3.6)	35.3 (3.1)
	A2	9.6 (1.5)	1931 (295)	4.95	41.6 (4.6)	35.7 (2.9)
	Test	9.6 (1.5)	1945(312)	4.95	45.7 (4.8)*	34.7 (2.8)
$noSR\text{-}noF_0$	A1	9.6 (1.5)	1913 (292)	4.95	43.2 (3.9)	35.1 (3.1)
	A2	9.6 (1.5)	1913 (292)	4.95	41.2 (4.2)	36 (2.8)
	Test	9.6 (1.5)	1944 (312)	4.95	45.7 (4.8)*	34.7 (2.8)
Greek						
$SR\text{-}F_0$	A1	8.1 (1.6)	1542 (278)	5.3 (0.8)	42.3 (5.1)	33.0(4.0)
	A2	7.9 (1.9)	1502 (302)	5.3 (0.6)	46.5 (9.2)	33.4 (3.3)
	Test	10.4 (1.5)*	1779 (255)*	5.9 (0.5)*	45.3 (6.1)	34.6 (2.8)
$SR\text{-}noF_0$	A1	8.5 (1.5)	1567 (288)	5.56 (0.6)	42.8 (4.9)	33.9 (3.4)
	A2	8.0 (1.9)	1507 (302)	5.3 (0.6)	46.9 (9.1)	33.3 (3.3)
	Test	10.4 (1.5)*	1779 (255)*	5.9 (0.5)*	45.3 (6.1)	34.6 (2.8)
$noSR\text{-}F_0$	A1	9.4 (1.8)	1689 (321)	5.55	44.0 (6.7)	35.5 (3.1)
	A2	9.4 (1.7)	1689 (308)	5.55	45.1 (8.3)	35.7 (3.47)
	Test	10.4 (1.5)*	1858 (264)*	5.55	45.3 (6.1)	34.6 (2.8)
$noSR\text{-}noF_0$	A1	9.1 (1.6)	1644 (293)	5.55	43.3 (6.3)	34.9 (3.6)
	A2	9.4 (2.0)	1689 (353)	5.55	46.3 (8.4)	35.9 (3.5)
	Test	10.4 (1.5)*	1858 (264)*	5.55	45.3 (6.1)	34.6 (2.8)
Korean						
$SR\text{-}F_0$	A1	10.3 (1.1)	2269 (239)	4.6 (0.7)	43.3 (7.8)	37 (2.3)
	A2	10.4 (1.6)	2242 (269)	4.7 (0.6)	41.2 (3.9)	36.8 (3.3)

	Test	11.6 (1.3)*	2323 (227)*	5.0 (0.4)*	53.9 (5.8)*	38.4 (2.9)
SR-noF$_0$	A1	10.3 (1.1)	2269 (239)	4.6 (0.7)	43.3 (7.8)	37.0 (2.3)
	A2	10.4 (1.6)	2242 (269)	4.7 (0.6)	41.2 (3.9)	36.8 (3.3)
	Test	11.6 (1.3)*	2323 (227)*	5.0 (0.4)*	53.9 (5.8)*	38.4 (2.9)
noSR-F$_0$	A1	11.3 (0.9)	2301 (190)	4.89	44.8 (6.3)	38.8 (3.6)
	A2	11.2 (0.8)	2288 (171)	4.89	42.8 (7.0)	38.4 (2.8)
	Test	11.6 (1.3)	2369 (275)*	4.89	53.9 (5.8)*	38.4 (2.9)
noSR-noF$_0$	A1	11.3 (0.9)	2301 (190)	4.89	41.4 (3.8)	38.4 (2.9)
	A2	11.2 (0.8)	2288 (171)	4.89	44.6 (7.7)	38.9 (3.4)
	Test	11.6 (1.3)	2369 (275)*	4.89	53.9 (5.8)*	38.4 (2.9)
Polish						
SR-F$_0$	A1	8.4 (1.5)	1643 (292)	5.2 (0.7)	44.1 (7.7)	33.6 (2.9)
	A2	8.7 (1.6)	1653 (242)	5.3 0.6)	43.7 (6.9)	34.0 (3.6)
	Test	11.8 (1.7)*	2084 (381)*	5.7 (0.6)*	40.9 (4.4)	37.8 (3.6)*
SR-noF$_0$	A1	8.2 (1.7)	1563 (226)	5.2 (0.9)	41.6 (6.1)	33.3 (3.4)
	A2	8.3 (1.9)	1596 (220)	5.2 (0.8)	41.1 (5.4)	33.6 (4.5)
	Test	11.8 (1.7)*	2084 (381)*	5.7 (0.6)*	40.9 (4.4)	37.8 (3.6)*
noSR-F$_0$	A1	8.9 (1.7)	1641 (316)	5.41	40.8 (4.4)	34.7 (3.3)
	A2	8.8 (2.4)	1617 (435)	5.41	43.5 (8.4)	35.4 (3.2)
	Test	11.8 (1.7)*	2180 (318)*	5.41	40.9 (4.4)	37.8 (3.6)*
noSR-noF$_0$	A1	9.1 (2.2)	1688 (404)	5.41	41.0 (3.6)	35.0 (3.6)
	A2	9.6 (1.1)	1779 (201)	5.41	41.6 (7.6)	35.7 (2.7)
	Test	11.9 (1.7)*	2180 (318)*	5.41	40.8 (3.7)	38.5 (3.3)*
Spanish						
SR-F$_0$	A1	9.1 (1.8)	1670 (288)	5.4 (0.5)	42.2 (6.1)	34.7 (4.4)
	A2	8.3 (1.5)	1590 (298)	5.3 (0.5)	42.2 (5.8)	34.6 (4.3)
	Test	11.4 (2.2)*	2049 (245)*	5.5 (0.6)	46.3 (4.3)*	36.5 (3.7)
SR-noF$_0$	A1	9.1 (1.8)	1670 (288)	5.4 (0.5)	42.2 (6.1)	34.7 (4.4)
	A2	8.3 (1.5)	1590 (298)	5.3 (0.5)	42.2 (5.8)	34.6 (4.3)
	Test	11.4 (2.2)*	2049 (245)*	5.5 (0.6)	46.3 (4.3)*	36.5 (3.7)
noSR-F$_0$	A1	10.1 (1.7)	1911 (322)	5.30	42.2 (5.7)	36.6 (3.5)
	A2	10.2 (1.8)	1923 (346)	5.30	43.1 (6.3)	36.8 (2.3)
	Test	11.4 (2.2)*	2147 (419)*	5.30	46.3 (4.3)*	36.5 (3.7)
noSR-noF$_0$	A1	10.3 (1.7)	1935 (320)	5.30	44.5 (7.3)	36.6 (3.6)
	A2	10.1 (1.9)	1911 (357)	5.30	42.3 (4.3)	37.0 (2.4)
	Test	11.4 (2.2)*	2147 (419)*	5.30	46.3 (4.3)*	36.5 (3.7)