Journal of Memory and Language 66 (2012) 665–679

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Language categorization by adults is based on sensitivity to durational cues, not rhythm class

Laurence White [a,*], Sven L. Mattys [b], Lukas Wiget [c]

[a] School of Psychology, Plymouth University, UK
[b] School of Experimental Psychology, University of Bristol, UK
[c] Department of General Linguistics, University of Zürich, Switzerland

ARTICLE INFO

Article history:
Received 17 November 2010
revision received 25 December 2011
Available online 3 March 2012

Keywords:
Speech perception
Speech production
Prosody
Rhythm class

ABSTRACT

Studies of listeners' ability to distinguish languages when segmental information is eliminated have been taken as evidence for categorical rhythmic distinctions between language groups (''rhythm classes''). Furthermore, it has been suggested that sensitivity to rhythm class is present at birth and that infants must establish the rhythm class of their native language as a precursor to language acquisition. We tested the hypothesis that adult listeners' ability to distinguish between languages is better predicted by differences in specific durational cues than by putative rhythm classes. We examined the categorization of language pairs using utterances in which only durational characteristics were preserved. We found that English listeners could distinguish between not only English and Spanish (from different rhythm classes), but also between different accents of British English. Furthermore, patterns of categorization between and within languages highlighted the contribution of speech rate, durational contrast and utterance-final lengthening.

Introduction

Adult listeners can distinguish pairs of languages like English and Spanish or Dutch and Japanese on the basis of prosodic information (Ramus, Dupoux, & Mehler, 2003; Ramus & Mehler, 1999). Language discrimination has also been demonstrated in human neonates (Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998) and older infants in the first year of life (Bosch & Sebastián-Galles, 1997). Such studies have been interpreted as indicating the existence of categorical prosodic distinctions between groups of languages (''rhythm classes'') that listeners are capable of perceiving and interpreting. Furthermore, the cognitive mechanisms used in rhythmic discrimination may be innate and fundamental to first language acquisition (Nazzi et al., 1998).

To focus on information in the speech signal that is relevant to ''rhythm class,'' segmental cues to language distinctions have traditionally been attenuated by low-pass filtering (e.g., Mehler et al., 1988) or removed by resynthesis (e.g., Ramus & Mehler, 1999). However, even such modified speech stimuli contain gradient variations, both between and *within* rhythm classes, in durational characteristics such as speech rate and utterance-final lengthening. Here, we tested the hypothesis that the perception of language differences is achieved through the exploitation of a range of fundamental prosodic cues rather than broad sensitivity to rhythm class.

Two components of rhythm in the auditory domain may be distinguished, ''coordinative rhythm,'' arising from the temporal organization of a string of sounds into a sequence of equally-timed (''isochronous'') groups, and ''contrastive rhythm,'' evident in any string of sounds in which there is an alternation of strong and weak elements. In speech, isochronous units, on the basis of which the different rhythm classes were originally proposed, have not been found (see Fletcher, 2010, for a summary), but contrastive rhythm is evident in the alternation between strong and weak syllables. Looking for evidence to support

* Corresponding author. Address: School of Psychology, University of Plymouth, UK.
E-mail address: laurence.white@plymouth.ac.uk (L. White).

0749-596X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.12.010

rhythm class distinctions on the basis of contrastive rhythm, Dauer (1983) identified phonetic and phonotactic regularities – in particular, shortening of vowels in unstressed syllables, and clustering of consonants in the onsets and codas of stressed syllables – that give rise to high durational contrast between stressed and unstressed syllables in languages, such as Dutch and English, that had been held to be "stress-timed" (i.e., to have equal time intervals between stressed syllables, Abercrombie, 1967). Languages with low durational stress contrast, such as French and Spanish, correspond to those held to be syllable-timed (i.e., to have equal syllable durations).

Approaches to quantifying contrastive speech rhythm have proposed metrics that exploit these phonetic and phonotactic patterns, measuring variation in vocalic and consonantal interval duration (e.g., Low, Grabe, & Nolan, 2000; Ramus, Nespor, & Mehler, 1999). Fig. 1 illustrates the distribution of scores, for several languages, of two of these metrics, VarcoV, the coefficient of variation of vocalic interval duration, and *%V*, the proportion of utterance duration comprised of vocalic intervals (see White & Mattys, 2007a). Languages like Dutch and English, particularly Standard Southern British English, have high VarcoV and low *%V* scores, due to the prevalence of consonant clusters and the large durational differences between stressed and unstressed vowels. In contrast, French and Spanish, with few consonant clusters and relatively low durational marking of stressed vowels, have low VarcoV and high *%V* scores.

There are, however, several arguments against the use of rhythm scores to map languages into rhythm classes such as "stress-timed" and "syllable-timed". First, the spread of scores within classes is at least as large as that between classes (e.g., Fig. 1). Second, some languages (e.g., Catalan) have relatively high stress contrast in vowel duration but simple consonant clusters, while others (e.g., Polish) show the opposite pattern, suggesting that they may be outside the standard rhythm classes (Nespor, 1990). Third, the rhythm class concept was based upon a hypothesis about isochrony of speech intervals (stress-delimited feet, syllables etc.) which, as mentioned earlier, has been demonstrated to be false.

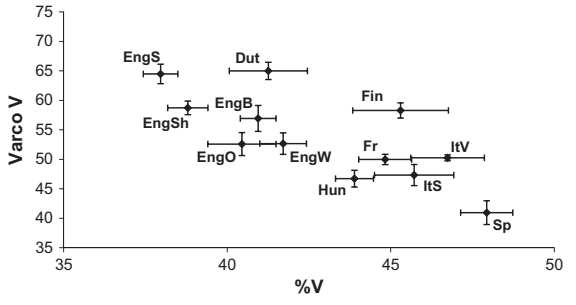

Fig. 1. Mean contrastive rhythm scores for a range of languages. VarcoV: coefficient of variation of vocalic interval duration. *%V*: vocalic proportion of total utterance duration. **Dut**: Dutch. **Eng**: English (**EngB**: Bristol; **EngO**: Orkney Islands; **EngS**: Standard Southern British; **EngSh**: Shetland Islands; **EngW**: Welsh Valleys). **Fin**: Northern Finnish. **Fr**: French. **Hun**: Hungarian. **It**: Italian (**ItV**: Veneto; **ItS**: Sicily). **Sp**: Spanish. Dutch, French, Hungarian, and Spanish speakers had (near-) standard (European) accents. *Original sources*: White and Mattys (2007a, 2007b) and White et al. (2009).

Given this lack of phonetic support for the notion of rhythm classes, the strongest evidence for the distinction comes from perceptual studies of language discrimination. Mehler et al. (1988) showed that neonates and 2-month-old infants could distinguish their own language from another language when listening to intact and even low-pass-filtered speech (see also Bosch & Sebastián-Galles, 1997; Dehaene-Lambertz & Houston, 1998). Mehler and Christophe (1995), reanalyzing Mehler et al.'s results, suggested that French neonates could also distinguish English from Italian. Critically, Nazzi et al. (1998) found that French neonates could discriminate low-pass-filtered English and Japanese, but not English and Dutch, and Christophe and Morton (1998) found that 2-month-old English-learning infants were not able to distinguish intact English utterances from Dutch utterances, although they discriminated both from Japanese.

While these studies suggest that infants can distinguish certain languages on the basis of prosodic cues, low-pass filtering merely attenuates rather than eliminates segmental information, and so infants may have attended to residual information about, for example, phonemic inventories or phonotactic regularities. To eliminate non-prosodic cues, Ramus and Mehler (1999) used resynthesis to generate speech in which the durations of consonantal and vocalic intervals were preserved, together with the pitch contour, while the phonemic inventory was simplified. Vowels were all replaced with /a/ and consonants were simplified by either replacement of a single consonant for each broad class (/s/ for fricatives; /t/ for stops etc. – the *saltanaj* condition), replacement of all consonants with /s/ (the *sasasa* condition), or replacement of all consonants and vowels with /a/ (one extended vowel with the original pitch contour, the *aaaa* condition). Ramus and Mehler found that adult French listeners' categorization of English and Japanese utterances was above chance in the *saltanaj* and *sasasa* conditions, but not in the *aaaa* condition. Finally, intonational variation was eliminated in a *flat sasasa* condition, but French listeners' ability to categorize English and Japanese utterances was still above chance. The fact that performance was no worse in the *flat sasasa* condition than the *saltanaj* and *sasasa* conditions was interpreted as evidence for the key role of speech rhythm.

Subsequently, Ramus et al. (2003) used *flat sasasa* utterances to compare discrimination of a range of language pairs, using an AAX same-different discrimination task rather than the categorization task of Ramus and Mehler. They found that adult French listeners' performance was predicted by rhythm class: "stress-timed" Dutch and English were not discriminated, but English was discriminated from "syllable-timed" Spanish and also from Catalan. This result is consistent with the rhythm class hypothesis, which predicts that listeners should fail to discriminate Dutch and English *flat sasasa* stimuli because the two languages are, in some fundamental sense, rhythmically equivalent. However, we suggest an alternative interpretation, namely, that listeners succeeded in this task when the differences between languages in terms of timing cues were large, regardless of rhythm class membership *per se* (see, e.g., scores in Fig. 1 for Dutch, English, and Spanish).

In addition to differences in the magnitude of rhythmic contrast, discrimination between languages in such studies may be facilitated by cross-linguistic differences in the distribution of stressed and unstressed syllables, in speech rate and in utterance-final lengthening. With regard to stress distribution, Spanish permits long sequences of unstressed syllables, whereas shorter sequences of unstressed syllables (one, two, or three) are the norm between stresses in English (Dauer, 1983). These distributional differences have no direct relationship to rhythm scores. With regard to rate, because Spanish has simpler consonant clusters than English, there will tend to be more syllables *per se*cond in Spanish than in English, everything else being equal. Finally, languages differ in their durational marking of prosodic structure: English, for example, has substantial utterance-final lengthening (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992), but this is attenuated in Spanish (see discussion in White and Mattys (2007a), White, Payne, and Mattys (2009) and Prieto, Vanrell, Astruc, Payne, and Post (2012)).

Listeners are capable of distinguishing phrases that differ in speech rate (Quené, 2007) and of perceiving utterance-final lengthening effects (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991), and so these cues may have contributed to previously-observed patterns of language discrimination. Moreover, where linguistic differences in these cues exist between languages within a rhythm class, such languages may be also distinguishable on the basis of durational information alone.

Purpose of the study

Given the range of potential cues, we wished to ascertain which ones listeners attend to when categorizing *flat sasasa* speech. According to the *class discrimination hypothesis*, only languages from different rhythm classes should be distinguished. This strong interpretation of previous studies is expressed by researchers such as Frota and Vigário (2001)[1] and Cho (2004).[2] Languages in different rhythm classes are likely to be characterized by differences in contrastive rhythm scores and in stress distribution. Languages within rhythm classes should not be distinguished, even if they have distinct rhythm scores.

Our alternative *durational contrast hypothesis* is that languages which differ substantially in contrastive rhythm scores should be distinguishable, regardless of membership of rhythm classes or patterns of stress distribution. To test this, we exploited differences in contrastive rhythm scores within English. Fig. 1 shows that accents of English from the Welsh Valleys (EngW) and the Orkney Islands (EngO) have rhythm scores intermediate between standard southern British English (EngS) and Castilian Spanish (Sp, see White & Mattys, 2007b, for interpretation of these scores). The class discrimination hypothesis predicts that these English accents should not be distinguishable on the basis of durational information. However, if performance is driven by the magnitude of differences in durational contrasts, rather than by a categorical distinction, EngS should be more easily distinguished from EngW or EngO than from Dutch.

We also considered two additional hypotheses, based on the potential contribution of speech rate and utterance-final lengthening to categorization performance. The *speech rate hypothesis* is that *flat sasasa* utterances should be categorized on the basis of differences in syllable-per-second speech rate. Under a strong version of this hypothesis, categorization should not be possible if rate differences are neutralized. The *final lengthening hypothesis* is that *sasasa* utterances are categorized on the basis of localized lengthening effects, specifically, the degree by which final syllables are longer than those in utterance-medial position.

Experiment 1

In Experiment 1, we attempted to provide an upper baseline for *flat sasasa* categorization by comparing two languages – EngS and Sp – from different "rhythm classes" and with large differences in contrastive rhythm scores, speech rate, and final lengthening. This experiment represents a replication, using an ABX rather than AAX paradigm, of one of Ramus et al.'s (2003) studies. Additionally, we attempted to determine the timing cues which were most predictive of categorization.

Method

Participants

We tested 24 native British English speakers, paid volunteers or University of Bristol undergraduates receiving course credit, with no self-reported speech or hearing problems. We did not control for the accent of British English spoken by participants, but most were speakers of near-Standard Southern British English. These criteria were the same in all experiments in this study.

Materials

The utterances on which the stimuli were based are shown in Appendix 1. The English sentences were constructed to be free of the approximants /l/, /r/, /j/, and /w/, in order to facilitate segmentation into vocalic and consonantal intervals. The Spanish sentences were likewise free of the approximants /l/, /r/, /j/ and /w/, although other allophonic approximants were not systematically excluded.

The English sentences were read by four speakers of Standard Southern British English and the Spanish

[1] "The scattering of languages along a rhythmic continuum does not seem to offer an explanation for the facts of perception. Several studies [...e.g., Mehler et al. (1988), Nazzi et al. (1998), Ramus and Mehler (1999)...have shown] that both adults and newborns, when exposed to filtered speech that preserves prosodic cues, are able to discriminate between languages belonging to different rhythm classes, but not between those of the same class. These findings strongly suggest that the properties behind the rhythmic distinctions are somehow encoded in the speech signal." (Frota & Vigário, 2001, p. 251)

[2] "According to their rhythmic properties, spoken languages have been classified by linguists into three categories, i.e. "stress-timed", "syllable-timed" and "mora-timed". Recently, this intuitive classification has been confirmed valid with perceptual studies. These studies have shown that both newborn infants and adults, upon hearing the pure rhythms of certain languages, are only able to discriminate between languages of the different categories mentioned above." (Cho, 2004, p. 249)

sentences were read by four speakers of Castilian Spanish. Speakers were instructed to read the sentences silently before reading them aloud, pausing between sentences, but trying not to pause within sentences. Recordings were made in sound-attenuated studios. These recordings and those for the subsequent experiments were a subset of the materials used in previous speech production studies (White & Mattys, 2007a, 2007b).

For each utterance, we measured the duration of all vocalic and consonantal intervals, identifying the boundaries between intervals based on visual inspection of the waveform and spectrogram, with occasional support from auditory evidence. Criteria for the identification of interval boundaries made reference to the characteristics of the onset/offset of vocalic formant structure and the shape of associated pitch periods in the waveform. Immediately-adjacent vowels were treated as a single interval. A consonantal interval was defined as beginning at the offset of a vocalic interval and ending at the onset of the next vocalic interval. Thus, as for vowels, immediately adjacent consonants were included in the same interval. For consistency, given that the goal was to generate a string of *sasasa* syllables, we excluded the first vocalic interval in the utterance and any preceding consonantal interval, so that the first measured interval was always the consonantal interval following the first vowel. For the same reason, any utterance-final consonantal intervals were also omitted from measurement. Pauses were excluded and, where intervals of the same type (vocalic or consonantal) were separated by a pause, these intervals were summed. These are standard procedures in the calculation of rhythm metrics (e.g., White & Mattys, 2007a), except for the exclusion of specific intervals required here to produce *sasasa* syllables.

The durations of vocalic and consonantal intervals, extracted from the labeled speech, were used to generate the *flat sasasa* stimuli for the perceptual experiment, following the procedure of Ramus and Mehler (1999), and to calculate contrastive rhythm scores and other durational parameters for all utterances. Each trimmed string of consonant and vowel intervals was used as input to the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van Der Vreken, 1996) to produce a sequence of *sasasa* syllables, each /s/ consonant and each /a/ vowel having the same duration as the corresponding interval in the original utterance. The fundamental frequency of the synthesized utterance was kept at a constant 230 Hz and there was no amplitude variation except for the contrast in intensity between /s/ and /a/.

We used the interval durations of the final *sasasa* utterances to derive the durational parameters, based on the formulae given in Table 1. As well as a set of standard contrastive rhythm metrics, we used three composite metrics of consonant + vowel interval duration (rPVI-CV, nPVI-CV, and VarcoCV). These were intended to capture variation in syllable duration, given that the subjective perceptual experience of *flat sasasa* speech is of a sequence of syllables rather than a sequence of consonants and vowels. We derived estimates of final lengthening (nFinal-C, nFinal-V, and nFinal-CV) for each utterance by dividing the duration of utterance-final intervals (/s/, /a/, and /sa/ combined) by the mean duration of the corresponding intervals. (This measure may underestimate the magnitude of the difference in final lengthening between English and Spanish, as all but one of the Spanish utterances ended in vowels, whereas the English utterances all ended in consonants. As final-lengthening tends to be progressive – i.e., greater when closer to the boundary – absolute-final vowels should show more final lengthening than those followed by consonants.)

Scores for all durational parameters are shown in Table 2. This table also shows the result of statistical tests, derived from mixed-effects linear regression models, comparing the means of durational parameters between the English and Spanish utterances. As can be seen, for Experiment 1, all English and Spanish utterances showed reliable or near-reliable differences on all durational parameters except MeanV and VarcoC.

Design and procedure

Following the protocol of Ramus et al. (2003), participants were told that they would hear modified speech from two mystery languages, which we termed "Sahatu" and "Eboda." Utterances were presented to participants, via headphones, using an ABX paradigm. On each trial, two *sasasa* utterances – the "example" utterances, A and B – were played consecutively, one from English and one from Spanish, followed by the test utterance, X, which could be from either language. Participants had to decide which of the example languages the X utterance belonged to (A or B), using the left shift key for language A and the right shift key for language B.

Within each ABX trial, the two utterances from the same language (e.g., A and X, or B and X) were always from different speakers and based on different sentences. Participants heard two blocks of twenty trials, with the same order of languages for the A and B example utterances throughout the two blocks. The order of presentation of the example languages (A and B) was counterbalanced between participants: thus, for 12 participants, Spanish was always the A example and English always the B example; for the other 12, English was always A and Spanish always B.

Within each block, each of the 20 utterances for each language was used as an example utterance (A or B) once. Half of the X utterances within the block were from each language, and each of the full set of 20 utterances for each language was used once as an X utterance, either in the first or the second block. Within these constraints, grouping within trials and ordering between trials were randomized, with the additional requirement that there could not be more than three A-correct or three B-correct trials in a row. Feedback – "correct" or "incorrect" – was given after each trial.

Statistical analysis

Two primary types of analysis are reported for this experiment and the following ones. Firstly, we took a signal detection approach to assess participants' discrimination of the two languages (Green & Swets, 1966). From each participant's hit rate and false alarm rate for each language, we calculated discrimination, d', which was

Table 1
Definitions of durational measures derived from *sasasa* utterances. For mathematical definitions of Pairwise Variability Index (PVI), see Grabe and Low (2002). Regarding the composite metrics (rPVI-CV, nPVI-CV and Varco-CV), we have elsewhere exploited metrics combining vowel + consonant, rather than consonant + vowel, sequences (Liss et al., 2009). The vowel + consonant sequence is generally a better acoustically-based approximation to the syllable, but incompatible with *sasasa* speech.

ΔV	Standard deviation of vocalic intervals
ΔC	Standard deviation of consonantal intervals
%V	Percentage of utterance duration comprised of vocalic intervals
nPVI-V	Normalized PVI for vocalic intervals. Mean of the differences between successive intervals divided by their sum (×100)
rPVI-C	PVI for consonantal intervals. Mean of the differences between successive intervals
MeanV	Mean duration of vocalic intervals
MeanC	Mean duration of consonantal intervals
VarcoV	Standard deviation of vocalic intervals divided by the mean (×100)
VarcoC	Standard deviation of consonantal intervals divided by the mean (×100)
rPVI-CV	PVI for consonant + vowel intervals. Mean of the differences between successive intervals
nPVI-CV	Normalized PVI for consonant + vowel intervals. Mean of the differences between successive intervals divided by their sum (×100)
Varco-CV	Standard deviation of consonant + vowel intervals divided by the mean (×100)
nFinal-C	Duration of final consonantal interval divided by the mean consonantal interval duration for the utterance
nFinal-V	Duration of final vocalic interval divided by the mean vocalic interval duration for the utterance
nFinal-CV	Duration of final consonant + vowel interval divided by the mean consonant + vowel interval duration for the utterance
Speech rate	Number of syllables per second

Table 2
Mean scores for the durational measures defined in Table 1 for the two languages in each experiment. The score for the first named language is always given first. *P*-values (derived from comparison of mixed-effects linear regression models) are indicated as: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$; $^{\dagger}p < .10$. For the rate-normalized experiments (2–5), MeanV and MeanC (in italics) were never used as predictive factors in the regression analyses, being wholly predictable from (and predictive of) *%V*.

	Experiment 1	Experiments 2–5		
	Sp vs EngS	Sp vs EngS	EngW vs EngS	EngW vs EngO
ΔV	34 vs 50***	38 vs 48**	44 vs 48*	44 vs 41
ΔC	41 vs 60***	46 vs 54**	46 vs 54***	46 vs 46
%V	48 vs 38***	49 vs 38***	44 vs 38***	44 vs 42†
nPVI-V	36 vs 71***	37 vs 73***	67 vs 73†	67 vs 74**
rPVI-C	44 vs 77***	50 vs 62***	50 vs 62***	50 vs 52
MeanV	80 vs 79	*92 vs 72****	*82 vs 72****	*82 vs 78*†
MeanC	89 vs 128***	*95 vs 115****	*105 vs 115****	*105 vs 109*†
VarcoV	42 vs 63***	41 vs 67***	53 vs 67***	53 vs 53
VarcoC	47 vs 47	48 vs 47	44 vs 47*	44 vs 43
rPVI-CV	51 vs 102***	58 vs 90***	79 vs 90*	79 vs 72
nPVI-CV	29 vs 49***	30 vs 47***	42 vs 47*	42 vs 39
Varco-CV	29 vs 42***	30 vs 42***	36 vs 42***	36 vs 33
nFinal-C	1.05 vs 1.19†	1.05 vs 1.21†	1.27 vs 1.21	1.27 vs 1.21
nFinal-V	1.15 vs 1.59*	1.12 vs 1.63**	1.65 vs 1.63	1.65 vs 1.26***
nFinal-CV	1.10 vs 1.35*	1.10 vs 1.37*	1.44 vs 1.37	1.44 vs 1.24**
Speech rate (syl/s)	6.0 v 4.9***	5.4 for all conditions		

compared to a null value of 0, and bias (*criterion*), *c*, compared to the null value of 0.5. *T*-tests were two-tailed in both cases.

Secondly, to identify the durational factors that listeners used to perform the discrimination task, we constructed a series of mixed-effect logistic regression models based on the raw response data – "correct" or "incorrect"– and including the random factors of subjects and trials (*lmer* package in *R*, Baayen, Davidson, & Bates, 2008). Models were compared using log-likelihood χ^2 tests.

Results and discussion

The mean percentage correct categorization was 26.3 out of 40 (66%). Discrimination was significantly better than chance, $d' = .87$, $t(23) = 7.59$, $p < .001$, and participants showed no response bias between languages, $c = .5$, $t(23) = .07$, $p > .10$. This result represents a replication of the result of Ramus and Mehler (2003) for English and Spanish, although their listeners were native French speakers, rather than the English speakers used here.

In our mixed-effect logistic regression models, we used as predictors the temporal parameters of the X utterances (see Table 1). Thus, in what follows, all factors – Language, Rate, ΔV, etc. – refer to the properties of the X utterances. Alternative predictive factors could have been explored, for example, based on temporal parameters for the A or B utterances, the difference between those utterances, the location of X on the A–B range, etc. However, it is the X utterance that is categorized and thus is presumably most salient at the point of decision. Furthermore, for the X utterance, the interpretation is clear (how strongly do X scores for the various temporal parameters reflect listeners' performance?), whereas alternative composite measures would have been less straightforward to interpret.

We fitted generalized linear mixed-effects regression models to the raw response data, starting with a basic model that only included the random factors of participants and items. A second model including Language (EngS vs Sp X utterance) showed no improvement over the initial model, indicating that the rate of correct categorization of *sasasa* X utterances was comparable in the two languages. In order to explore additional predictive factors, we selected all of the durational parameters that showed (near-) significant differences between English and Spanish (i.e., all parameters except for MeanV and VarcoC, see Table 2). For each such parameter, we then constructed two further models, building upon the two models described above. The third model included the main effect

of Language, as before, and the main effect of the parameter in question (ΔV, ΔC, etc.). The fourth model additionally included the interaction of that parameter with Language, on the basis that the direction of prediction should be different for English and Spanish (e.g., high ΔV X utterances would be expected to be well categorized for English and low ΔV X utterances for Spanish). We identified all the models which showed a main effect of Language or an interaction between Language and the parameter. We then used Akaike's information criterion (AIC, Akaike, 1974) to identify the best-fitting of those models (i.e., the model with the lowest AIC).

We found that the best-fitting model was the one that included an interaction between Language and Speech Rate. This interaction, χ^2 (1) = 27.69, $p < .001$, reflected the slower rate of the English than Spanish utterances (Table 2, EngS: 4.9 syl/s vs Sp: 6.0 syl/s). To explore this interaction further, we fitted separate regression models for the X utterances of the two languages. For both English X utterances and Spanish X utterances, models including speech rate showed substantial improvement over a simple random effects model: EngS, $\chi^2(1) = 20.12$, $p < .001$; Sp, $\chi^2(1) = 10.39$, $p < .01$. Furthermore, these were the best predictive models for each language. The addition of further durational parameters as factors did not significantly improve the models.

As shown in Fig. 2, the Spanish X utterances were more likely to be correctly categorized if they had higher speech rate, whereas the reverse was true for the English utterances. Despite the reliable differences on scores for many other durational parameters between the English and Spanish utterances (Table 2), the addition of these parameters to the models that already included Rate conferred no further predictive benefit.

The effect of speech rate on language discrimination has not been directly examined in previous studies. As discussed above, the phonotactics of Spanish and English mean that consonantal interval durations tend to be shorter for Spanish. There will consequently tend to be more syllables per second in Spanish than in English, even if mean vowel durations are comparable. These durational trends were indeed observed in our utterances: mean vowel duration: EngS 79 ms vs Sp 80 ms; mean consonant duration: EngS 128 ms vs Sp 89 ms. Thus, although the number of syllables per utterance was controlled across the two languages (EngS 14.4 vs Sp 14.8, $p > .10$), there was a significant difference in mean utterance duration (EngS 2968 ms vs Sp 2495 ms, $p < .0001$), and hence, speech rate (see also Dellwo, 2008, for a report of systematic differences between languages in speech rate). Although it is plausible that participants could use utterance duration instead of, or in addition to, speech rate as the primary cue, this possibility is unlikely because Ramus and Mehler (1999) showed that discrimination of stimuli that differed only in total duration and fundamental frequency was not above chance.

Finally, we examined the effect of feedback. The rationale for giving feedback ("correct" or "incorrect") after each trial was to maintain participants' concentration over a difficult and repetitive task, but it is possible that such feedback may also promote progressive learning of the

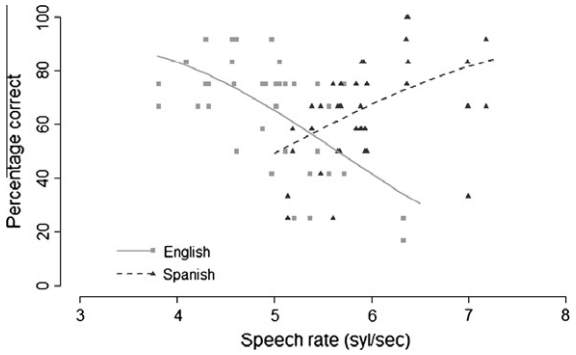

Fig. 2. Speech rate and percentage correct categorization for Spanish and English X utterances in Experiment 1. Logistic regression lines, indicating the relationships between correct categorization and speech rate, are derived from the raw binomial data.

salient cues to categorization. If so, performance should increase over the course of the experiment. However, including trial order as a factor did not improve the basic logistic regression model, $\chi^2(1) = 0.40$, $p > .10$, indicating that discrimination performance was not being driven by learning during the experiment.

To further and more directly ascertain that feedback had no substantial effect on our results, we re-ran Experiment 1 without trial-by-trial feedback. Mean percentage correct categorization was 23.1 out of 40 (58%), and discrimination was significantly better than chance, $d' = .41$, $t(22) = 3.48$, $p < .005$. There was no response bias between languages, $c = .5$, $t(22) = .38$, $p > .10$, and no effect of trial order, $\chi^2(1) = 1.36$, $p > .10$. However, performance was lower than in the same experiment run with feedback, $\chi^2(1) = 13.32$, $p < .001$. Given the lack of evidence for progressive learning in the order, this feedback vs no-feedback difference suggests that feedback facilitates concentration during the task rather than learning per se. In line with previous studies (Ramus & Mehler, 1999; Ramus et al., 2003), the feedback element was retained for the other experiments.

In sum, Experiment 1 replicates the finding of Ramus et al. (2003), namely, that English and Spanish can be distinguished on the basis of purely durational information, and points to the importance of speech rate as a cue to categorization: As in our experiment, rate was not systematically controlled in the Ramus et al. study. This result supports the speech rate hypothesis outlined above. The strong version of this hypothesis holds that rate differences are necessary for categorization. Thus, we next considered whether the same two languages can be distinguished when speech rate is neutralized.

Experiment 2

In Experiment 2, we examined whether English listeners could perform the ABX categorization task on English and Spanish *sasasa* utterances once speech rate differences were neutralized. Differences in temporal stress contrast and other durational parameters listed in Table 2 were maintained, allowing examination of the contribution of

these parameters to language categorization in the absence of rate differences.

Method

Participants and materials

We tested 24 native British English speakers. In order to remove all rate information from the utterances used in Experiment 1, we first truncated the sequence of interval durations for each utterance to leave just twelve syllables in total, removing as many syllables as necessary from the start of the utterance. Then we expanded or compressed the interval durations of each utterance uniformly to generate utterances of 2240 ms total duration, maintaining the relative durations of vowels and consonants from the original utterances. These interval durations were used, as before, as input to the MBROLA synthesizer. Thus, in addition to the absence of segmental and pitch variation, the rate-normalized *sasasa* sentences were equalized in terms of syllable number and overall syllable-per-second speech rate.

The procedure for the calculation of durational measures was the same as for Experiment 1. Scores for all measures and statistical comparisons of means between the English and Spanish utterances are shown in Table 2. As can be seen, utterances showed a comparable pattern of differences to those in Experiment 1, except that, as a result of normalization, languages no longer differed in speech rate and a difference in mean vowel duration emerged.

Design and procedure

The design, procedure and statistical analyses were identical to those in Experiment 1.

Results and discussion

The mean percentage correct categorization was 22.5 out of 40 (56%), and discrimination was significantly better than chance, $d' = .38$, $t(23) = 2.67$, $p < .05$. There was no response bias between languages, $c = .47$, $t(23) = -1.23$, $p > .10$, and no effect of trial order, $\chi^2(1) = 2.12$, $p > .10$. However, there was a significant drop in categorization performance compared to Experiment 1, $\chi^2(1) = 18.19$, $p < .001$, which confirms the importance of speech rate. The fact that performance was still above chance suggests that listeners were nevertheless able to exploit other cues.

As in Experiment 1, we attempted to determine which temporal properties of the X utterances were the most predictive of categorization performance. We selected those durational parameters that showed significant differences between rate-normalized English and Spanish utterances (Table 2) and ranked them according to AIC obtained from logistic regression models for both languages combined, with Language and each durational parameter in turn as factors. These factors were used in a stepwise fashion, according to ranking, in separate logistic regression models for English and Spanish, until no further factors could be added to improve the models. The same statistical procedure was followed for all subsequent experiments.

There was no effect of Language, and thus no difference in the correct categorization rate of EngS and Sp X utterances. For the English X utterances, the durational factor which best predicted categorization was nFinal-V: $\chi^2(1) = 5.28$, $p < .05$. This factor is intended to measure the degree of lengthening of the utterance-final vowel, being, as described in Table 1, the duration of the final vowel divided by the mean vowel duration of the whole utterance. As shown in Table 2, there was much greater final lengthening in English utterances than in Spanish utterances (nFinal-V: EngS 1.63 vs Sp 1.12). This analysis suggests that English listeners used the degree of final lengthening of X utterances as a cue to language categorization. This model was further improved by adding rPVI-C (the pairwise variability index of consonant duration, see Tables 1 and 2), $\chi^2(1) = 4.40$, $p < .05$. This two-factor model was not further improved by the addition of any other factors. The contribution of rPVI-C suggests that the contrast in duration of /s/-onsets between successive *sa* syllables was a useful cue for listeners. This clearly relates to timing differences between English and Spanish, the latter having fewer and simpler consonant clusters, hence shorter consonant intervals with less variation in the duration of successive intervals.

For the Spanish X utterances, rPVI-C was the only significant predictor of categorization performance, $\chi^2(1) = 5.76$, $p < .05$. Thus, variation in duration between successive syllable onsets was useful for categorization of both Spanish utterances (low contrast) and English utterances (high contrast). This symmetry between languages did not apply to the final lengthening cue: nFinal-V was not a predictive factor for Spanish X utterances, but listeners were more likely to categorize English X utterances correctly when there was a large degree of final-vowel lengthening. We return to this point in the discussion of Experiment 4.

The results of Experiment 2 demonstrate that listeners can distinguish utterances from English and Spanish on the basis of purely durational cues, even in the absence of speech rate variation. Thus, the strong version of the speech rate hypothesis must be rejected, but the drop in performance from Experiment 1 to Experiment 2 supports a weaker version of the hypothesis (i.e., speech rate is *one* of the cues to language categorization). The nature of the predictive factors – nFinal-V and rPVI-C – provides support for both the final lengthening hypothesis and the durational contrast hypothesis. However, in order to clearly distinguish between these hypotheses and the class discrimination hypothesis, we need to see whether discrimination is possible within a "rhythm class." This is the purpose of Experiment 3 and Experiment 4.

Experiment 3

In Experiment 3, we examined whether listeners can distinguish between *flat sasasa* utterances taken from languages within the same "rhythm class." We contrasted utterances of Standard Southern British English (EngS) and Welsh Valleys English (EngW). These are two varieties previously established to differ on metrics of temporal stress contrast such as VarcoV and %*V* (White & Mattys, 2007b). As noted above, the differences in rhythm scores between these varieties of English are substantially larger than those between EngS and Dutch (Fig. 1), two languages previously used in within-rhythm-class discrimination experiments (e.g., Ramus et al., 2003). These two English varieties have the additional advantage that differences

in temporal stress contrast are not paralleled by differences in utterance-final lengthening (in contrast with the situation for EngS vs Sp), so this comparison also allowed us to eliminate final lengthening, shown to be a useful cue in Experiment 2.

Furthermore, the use of a single set of sentences for the EngS vs EngW contrast allowed us to eliminate differences in stress distribution and focus on cues relating to variation in interval duration, as indexed by metrics of temporal stress contrast.

Method

Participants and materials

We tested 24 native English speakers. The EngS *sasasa* utterances were those used in Experiment 2. The EngW utterances were based on recordings of four speakers of Welsh Valleys English reading the same set of five sentences used for the EngS utterances. The procedure for making the interval duration measurements was the same as for the previous experiments. The production of experimental utterances, using MBROLA synthesis, followed the rate-normalization procedure used in Experiment 2 (the EngS sentences were re-used from that experiment). As can be seen in Table 2, EngS and EngW showed significant differences on most measures of variation in duration of vocalic and consonantal intervals, although these differences were generally smaller than the differences between EngS and Sp. There were no significant differences in any of the final lengthening measures, however, with scores for both EngS and EngW indicating substantial degrees of final lengthening.

Design and procedure

The design, procedure, and method of statistical analysis were identical to those of Experiment 1.

Results and discussion

The mean percentage correct categorization was 22.3 out of 40 (56%) and discrimination was significantly better than chance, $d' = .30$, $t(23) = 3.54$, $p < .005$. There was no response bias between languages, $c = .48$, $t(23) = -1.58$, $p > .10$, and no effect of trial order, $\chi^2(1) = 0.01$, $p > .10$. Categorization performance in Experiment 3 was worse than in Experiment 1, $\chi^2(1) = 20.63$, $p < .001$, but there was no difference between Experiments 2 and 3.

As described above, we used mixed-effects logistic regression models to determine the factors that best predicted categorization performance. There was no main effect of Language, indicating no difference in correct categorization between EngS and EngW X utterances.

For the EngS X utterances, the durational factor that best predicted categorization was ΔC, the standard deviation of consonantal interval duration, $\chi^2(1) = 5.25$, $p < .05$ (see Tables 1 and 2). Thus, as in Experiment 2, a metric of consonantal interval variability (ΔC) was found to be predictive of categorization performance. In Experiment 2, however, the metric was rPVI-C. A possible reason for the emergence of different predictors in the two experiments is that PVI metrics may be affected by stress distribution, which – as discussed above – differs between Spanish and English (Experiment 2). However, there were no overall stress distribution differences between the two sets of English utterances (EngS and EngW, Experiment 3) and so a PVI-based metric would not be expected to be more effective than a metric of overall variability in consonantal interval duration. The next best predictor was ΔV, the standard deviation of vocalic interval duration. Its contribution to the best-fitting model, in addition to ΔC, approached significance: $\chi^2(1) = 3.27$, $p < .10$.

For the EngW X utterances, ΔC was again the factor that best predicted categorization, $\chi^2(1) = 4.06$, $p < .05$. Thus, utterances with low variation in consonantal interval duration tended to be categorized as EngW and utterances with high variation as EngS. The inclusion of a further factor, *%V*, in the best-fitting model approached significance, $\chi^2(1) = 2.73$, $p < .10$. As noted above, *%V* (see Tables 1 and 2) was perfectly correlated with MeanV and MeanC for these rate-normalized utterances: knowing the value of one parameter determines the other two, so the nature of the cue indicated by this effect is not clear. However, the information conveyed by such a cue must derive from a global rather than a local property of the utterance: mean vowel duration, mean consonant duration, or the balance between the two.

Experiment 3 demonstrates that language categorization on the basis of purely durational information is possible within a "rhythm class," contrary to the class discrimination hypothesis. Specifically, English listeners could distinguish between two varieties of their own language – EngS and EngW – that differed on metrics of temporal stress contrast. This finding supports the durational contrast hypothesis. It also suggests that, although useful (cf. rPVI-C in Experiment 2), a contrast in stress *distribution* is not necessary for categorization. Likewise, despite the predictive power of nFinal-V found in Experiment 2, categorization was achieved here between two varieties that lacked a difference in final lengthening. In Experiment 4, we again compared two accents of English, but we reversed the cue availability, neutralizing differences of temporal stress contrast to focus on the use of final lengthening.

Experiment 4

Having demonstrated that duration-based categorization is possible within a "rhythm class," in Experiment 4 we examined whether within-class categorization is possible even where there are minimal differences in temporal stress contrast. To this end, participants performed the categorization task on two further accents of English: Orkney English (EngO) and Welsh Valleys English (EngW). As shown in Fig. 2 and Table 2, these accents are similar in terms of metrics of temporal stress contrast (see White & Mattys, 2007b, for interpretation of these patterns). Furthermore, as in Experiment 3, "rhythm class" differences and stress distribution differences are neutralized. However, the two accents have a clear difference in final lengthening (Table 2): compared to EngW, EngO has attenuated lengthening of the utterance-final vowel and, consequently, of the utterance-final syllable. Thus, this experiment is a strong test of whether categorization is

possible on the basis of differences in the magnitude of a localized, utterance-final, lengthening effect.

Method

Participants and materials

This experiment included 24 native English speakers. The EngW *sasasa* utterances were those of Experiment 3. The Orkney English utterances were based on recordings of four speakers of EngO reading the same set of five sentences as used for the EngS and EngW utterances. The procedure for constructing the *flat sasasa* utterances was the same as in the previous experiments. Durational measures are shown in Table 2.

Design and procedure

These were the same as in the previous experiments.

Results and discussion

The mean percentage correct categorization in the comparison between EngW and EngO was 21.2 out of 40 (53%), and discrimination was significantly better than chance, $d' = .17$, $t(23) = 2.21$, $p < .05$. There was no response bias between languages, $c = .52$, $t(23) = -.78$, $p > .10$. Categorization performance in Experiment 4 was worse than in Experiment 1, $\chi^2(1) = 32.70$, $p < .001$, but not significantly different from Experiment 2 or Experiment 3 ($ps > .10$). Unlike in the earlier experiments, there was an effect of trial order, $\chi^2(1) = 3.93$, $p < .05$. Inspection of the means for ordered sets of ten trials reveals that performance deteriorated over the course of the experiment – Set 1: 56%; Set 2: 55%; Set 3: 53%; Set 4: 48% – suggestive of a fatigue effect in this difficult task.

We attempted, as before, to determine the predictive durational factors from the limited set of parameters on which EngW and EngO utterances differed (Table 2). There was no effect of Language, indicating no difference in the rate of correct categorization of EngO or EngW X utterances. The final-syllable lengthening parameter, nFinal-CV, was predictive of categorization performance for the EngW X utterances, $\chi^2(1) = 8.72$, $p < .01$. Thus, as in Experiment 2, listeners categorized X utterances with substantially lengthened final syllables as belonging to the EngW group. None of the factors listed in Table 2 predicted categorization performance for the EngO X utterances. Again, as in Experiment 2, it seems that English listeners did not use the absence of substantial final lengthening to categorize utterances as belonging to EngO. This suggests a bias, at least for English listeners, to pay attention to large degrees of utterance-final lengthening. It should be noted that, although there was a significant difference between EngO and EngW in nPVI-V scores, and a near-significant difference in %*V* scores, neither of these factors were found to be predictive of the categorization performance.

In summary, the final lengthening hypothesis is supported once again, showing that listeners can use substantial lengthening of the utterance-final syllable as a cue for categorization, even when contrastive rhythm differences are minimal. Thus, together with the previous experiments, Experiment 4 demonstrates that listeners can categorize utterances on the basis of durational information.

The comparison between the correct categorization rates of Experiment 1 and the other experiments makes it clear that speech rate differences are utilized in preference to other cues and that categorization is adversely affected in the absence of rate differences (Table 3). We also compared categorization performance in Experiments 2–4, but, as summarized in Table 3, we did not find any significant differences. This could indicate that the other available cues – stress distribution, temporal stress contrast, utterance-final lengthening – were equivalent in their effectiveness, singly or in combination. However, comparisons between these experiments were all between subjects, and so may have lacked the power to reveal potentially subtle differences. In Experiment 5, we tested the strength of different cues more stringently by repeating Experiments 2–4 in design that allows for comparison of performance within subjects.

Experiment 5

There are several reasons to expect that categorization within languages (in Experiments 3 and 4) should be more difficult than between languages (Experiment 2). First, the stress distribution contrast that exists between English and Spanish – more regular alternation of stressed and

Table 3
Comparisons of categorization performance between experiments. Differences were evaluated using a log-likelihood χ^2 test, comparing mixed-effects logistic regression models with and without Experiment as a factor. Statistically reliable differences are shown in bold. For all significant differences, the condition (underlined) on the left had the higher performance.

		Rate normalized	
	Sp vs EngS	EngW vs EngS	EngW vs EngO
Between subjects comparisons			
Experiment 1 (Sp vs EngS, full rate) vs Experiments 2–4	**$\chi^2(1) = 18.19$, $p < .001$**	**$\chi^2(1) = 20.63$, $p < .001$**	**$\chi^2(1) = 32.70$, $p < .001$**
Experiment 1 (Sp vs EngS, full rate) vs Experiment 5	**$\chi^2(1) = 7.11$, $p < .01$**	**$\chi^2(1) = 27.99$, $p < .001$**	**$\chi^2(1) = 27.27$, $p < .001$**
Experiment 2 (Sp vs EngS, rate-norm) vs Experiments 3 and 4		$\chi^2(1) = 0.05$, $p > .10$	$\chi^2(1) = 2.15$, $p > .10$
Experiment 3 (EngW vs EngS, rate-norm) vs Experiment 4			$\chi^2(1) = 1.54$, $p > .10$
Within subjects: all comparisons are between Experiment 5 conditions (all rate-normalized)			
Condition A (Sp vs EngS) vs Conditions B and C		**$\chi^2(1) = 7.00$, $p < .01$**	**$\chi^2(1) = 6.62$, $p < .05$**
Condition B (EngW vs EngS) vs Condition C			$\chi^2(1) = 0.01$, $p > .10$

unstressed syllables in the latter – is absent in the within-language conditions, where speakers of all varieties of English are reading the same texts. Second, in the EngS vs Sp comparison, listeners can exploit both differences in temporal stress contrast and differences in final lengthening, whereas the within-English comparisons only provide one of these cues, not both: only temporal stress contrast for EngS vs EngW; only final lengthening for EngO vs EngW. Third, the magnitude of the within-language differences in these durational cues – temporal stress contrast and final lengthening – was smaller than the between-languages differences (see Table 2). Experiment 5 allowed us to compare these conditions directly.

Method

Participants and materials

This experiment included 24 native English speakers. The experimental stimuli were those used in Experiments 2–4. All of them were rate-normalized.

Design and procedure

All 24 participants completed Experiments 2–4 in a single three-condition design: Condition A, EngS vs Sp (replication of Experiment 2); Condition B, EngS vs EngW (replication of Experiment 3); Condition C, EngO vs EngW (replication of Experiment 4). For each participant, the three conditions were carried out on three separate days, with the order of experiments fully counterbalanced between participants. The four "mystery languages" were variously given the names "Sahatu," "Eboda," "Moltec," and "Ventish."

In this experiment, we offered monetary prizes for the participants who scored most highly over the three conditions, because, having established a baseline for performance in Experiments 2–4, we wished to assess the optimal level of performance possible given the cues available. As the results show, monetary incentive turned out to have minimal effect on performance levels compared with the previous experiments. As in the other experiments, participants received feedback – "correct" or "incorrect" – after each trial.

Results and discussion

Above-chance categorization of the X utterances was observed for all three conditions: Condition A (EngS vs Sp): 24 out of 40 (60%), $d' = .58$, $t(23) = 3.25$, $p < .01$; Condition B (EngS vs EngW): 21.6 out of 40 (54%), $d' = .21$, $t(23) = 3.35$, $p < .01$; Condition C (EngO vs EngW): 21.7 out of 40 (54%), $d' = .25$, $t(23) = 2.21$, $p < .05$. There was no significant response bias between languages for Condition A, $c = .48$, $t(23) = -.78$, $p > .10$, and Condition B, $c = .47$, $t(23) = -1.38$, $p > .10$. In Condition C, participants showed a near-significant bias to select EngO X utterances, $c = .54$, $t(23) = 2.04$, $p = .053$, a bias which underpins the between-language difference in correct scores for this experiment (see below). There were no effects of trial order.

Categorization rates for this experiment and Experiments 1–4 are shown in Fig. 3. For EngS vs EngW and for EngO vs EngW, performance was not statistically different in the two versions of the experiments ($ps > .10$). In the EngS vs Sp comparison, there was a slight trend towards higher categorization performance in Condition 5A than in Experiment 2, $\chi^2(1) = 2.72$, $p = .10$.

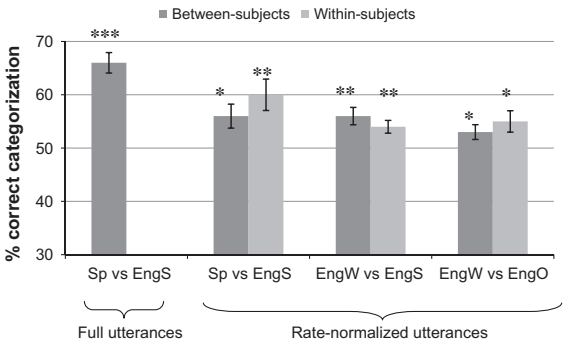

Fig. 3. Mean percentage correct categorization and standard-error bars for each experiment. *P* values, derived from the comparison of d' values with a null value of zero, are indicated thus: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

Our prediction regarding relative performance was confirmed (Table 3): Categorization was more effective when the two sets of samples were from different languages (English and Spanish) than when they were from two varieties of the same language (EngS vs EngW; EngO vs EngW). As discussed above, the performance drop from Condition A to Conditions B and C may relate to: smaller/absent differences in contrastive rhythm scores; smaller/absent differences in final lengthening; absence of stress distribution differences. The lack of difference between Conditions B and C suggests that neither interval contrast nor utterance-final lengthening strongly dominated the other cue, but performance was better when both were present than when only one was available.

To further explore the relative contribution to categorization of these factors, we applied the predictive models used in Experiments 2–4 to the data of Experiment 5. The best predictors for categorization of each language in each experiment are summarized in Table 4, including the results of all relevant statistical tests. Where the factors for these two sets differ, this indicates that the factors identified as predictive for Experiments 2–4 were not

Table 4
X utterance durational parameters that were predictive of correct categorization performance in the rate-normalized experiments. Predictors were assessed in mixed-effect logistic regression models and ranked according to Akaike's Information Criterion. *P*-values are indicated as: $^{***}p < .001$; $^{*}p < .05$; $^{\dagger}p < .10$.

	Experiments 2, 3, 5	Experiment 5
Sp vs EngS		
Sp	rPVI-C: $\chi^2(1) = 5.76^{*}$	ΔV: $\chi^2(1) = 5.51^{*}$
EngS	nFinal-CV: $\chi^2(1) = 5.28^{*}$	nFinal-CV $\chi^2(1) = 13.86^{***}$
	rPVI-C: $\chi^2(1) = 4.40^{*}$	ΔV: $\chi^2(1) = 3.21^{\dagger}$
EngW vs EngS		
EngW	ΔC: $\chi^2(1) = 4.06^{*}$	ΔC: $\chi^2(1) = 5.29^{*}$
	$\%V$: $\chi^2(1) = 2.73^{\dagger}$	
EngS	ΔC: $\chi^2(1) = 5.25^{*}$	*No predictive factors*
	ΔV: $\chi^2(1) = 3.27^{\dagger}$	
EngW vs EngO		
EngW	nFinal-CV: $\chi^2(1) = 8.72^{**}$	nFinal-V: $\chi^2(1) = 3.05^{\dagger}$
EngO	*No predictive factors*	*No predictive factors*

predictive for Experiment 5, and so we attempted to find alternative combinations of factors.

There was no difference between the correct categorization rates for EngS and Sp X utterances (Condition A). For the EngS X utterances, we found that nFinal-V was predictive of categorization, as for Experiment 2. The other predictive factor, for both EngS and Sp X utterances, was ΔV (see Table 4), in contrast with Experiment 2 in which rPVI-C was predictive. However, in both cases, participants were attending to a combination of local and global cues (final lengthening and interval duration contrasts, respectively). As categorization performance was better for EngS vs Sp than for the two other comparisons, this suggests that the combination of cues produced better performance than either in isolation.

For the EngS and EngW comparison (Condition B), there was a near-significant effect of Language, $\chi^2(1) = 3.02$, $p = .08$, which indicates that the difference in categorization rates is probably robust, despite the absence of a bias in the signal detection analysis, EngS: 51% vs EngW: 57% (this difference was not observed in Experiment 3). Indeed, the rate of correct categorization was not significantly above chance for EngS, and there were no predictive factors of performance for EngS X utterances. However, ΔC was predictive of categorization for the EngW X utterances, as in Experiment 3. English listeners were thus sensitive to variation in consonant duration, with X utterances that had relatively low variation being more likely to be correctly categorized as EngW.

For the EngO vs EngW comparison (Condition C), there was a main effect of Language, $\chi^2(1) = 5.50$, $p < .05$, indicating that performance was significantly higher for the EngO than EngW utterances (58% vs 50%, respectively, a difference not observed in Experiment 4). As shown in Table 4, a metric of final lengthening was predictive of categorization performance for the EngW X utterances, but this cue did not result in an overall above-chance result. This caveat did not apply in Experiment 4, where final lengthening was a clear cue to categorization: EngW X utterances with relatively long final syllables were more likely to be categorized into the correct group. As in the EngS vs Sp comparison, English listeners were sensitive to the presence, but not the absence, of substantial final lengthening.

With regard to the emergence of language effects in two conditions of Experiment 5 where there were none in the earlier experiments, we offer the suggestion that exposure to multiple languages in the same design may promote confusion in participants regarding the best cues for each one. They may resolve this strategically by focusing on identification of one language rather than both.

Overall, the results of Experiment 5 show that listeners performed better on the between-language comparison (EngS vs Sp) than on the within-language comparisons (EngS vs EngW, EngO vs EngW). The use of both final lengthening and durational contrast whenever available suggests that performance fell because only one of these two cues was available in each of the within-languages comparisons. The greater magnitude of timing differences between languages than within, and the lack of stress distribution differences within languages may also have contributed to the performance drop.

General discussion

The experiments reported here demonstrate that adult listeners can distinguish languages purely on the basis of durational information (cf. Ramus & Mehler, 1999). We exploited natural variability between and within languages to distinguish several hypotheses regarding the specific durational cues which listeners use for this purpose.

Speech rate hypothesis

The speech rate hypothesis was that utterances are categorized on the basis of differences in syllables-per-second speech rate. This hypothesis was strongly supported by the result of Experiment 1, where speech rate was the only predictive factor for correct categorization of EngS and Spanish utterances, and by the substantial drop in performance when rate was normalized (Experiment 2). We conclude that not only do listeners use variation in the rate of occurrence of syllables to distinguish between languages, but that this is the primary cue exploited where available (see also Dellwo, 2008, for a discussion of the importance of rate in the perception of speech rhythm).

The current results suggest that perception of rate differences probably contributed to patterns of between-rhythm-class discrimination found in previous studies. Indeed, our EngS and Sp samples had a mean rate difference of 22%, which is well outside the 5% just noticeable difference in rate found by Quené (2007). Furthermore, according to the characterization by Dauer (1983) of the stress-timed vs syllable-timed distinction, languages from different classes, like English and Spanish, will always tend to differ in syllables-per-second rate. "Mora-timed" Japanese, with simpler syllable structures than Spanish, will tend to show an even greater rate difference from "stress-timed" languages like Dutch, with which it has been compared in previous discrimination experiments (e.g., Ramus & Mehler, 1999). Thus, between-rhythm-class discrimination of modified speech may often be reducible to rate discrimination in the absence of explicit control of rate.

However, our rate-normalized comparisons demonstrate that listeners can categorize utterances with some accuracy even when differences in syllables-per-second rate are eliminated. Thus, while rate is clearly a factor in language categorization, we must reject the strongest version of the speech rate hypothesis, that categorization is not possible if rate differences are neutralized. Listeners evidently rely on other cues in the absence of rate differences.

Durational contrast hypothesis

The durational contrast hypothesis was that languages which differ in contrastive rhythm scores should be distinguishable, regardless of "rhythm class" membership or patterns of stress distribution. Experiments 2, 3, and 5 showed that, when speech rate differences were neutralized, several metrics were predictive of categorization: metrics of consonant duration (rPVI-C, ΔC), vowel duration

(ΔV), and the durational balance between consonants and vowels (*%V*).

With regard to consonantal metrics, rPVI-C was predictive of categorization in the comparison between languages, EngS vs Sp (Experiment 2), but ΔC was a better predictor for the within-language comparison, EngS vs EngW (in both Experiment 3 and Condition 5b). PVI metrics are intended to reflect the within-utterance pattern of alternation between long and short intervals, information which may be salient where differences exist in stress distribution. For example, the contrast between long and short consonantal intervals should occur with greater frequency in English than in Spanish, and the superiority of rPVI-C over ΔC for this comparison suggests that listeners exploit this distributional difference, as well as the overall degree of consonantal variability. For vocalic variation, however, the standard deviation of vowel duration was predictive rather than the PVI metric, even where differences in stress distribution were present (i.e., between English and Spanish). Thus, listeners utilized global rather than local information about vowel duration. The reasons for this strategic difference in the exploitation of consonantal and vocalic variation are not clear.

Listeners apparently attended to sub-syllabic rather than syllabic durational variation, as none of the composite metrics of variation in *sa* duration were found to be predictive. The predictive power of *%V*, found in one comparison here and by Ramus et al. (1999), also reflects the perceptual separation between vowels and consonants, although it is not clear whether the critical factor is mean vowel duration, mean consonant duration, or the balance between the two.

Finally, we found that contrastive rhythm metrics were, at best, subsidiary predictors when large final lengthening effects were available as categorization cues.

Final lengthening hypothesis

The final lengthening hypothesis was that utterances can be categorized on the basis of localized lengthening effects, specifically, the degree to which final syllables are longer than those in utterance-medial position. This hypothesis was clearly supported in the two rate-normalized comparisons for which the languages differed significantly in the degree of utterance-final lengthening: EngS vs Spanish; EngO vs EngW. However, there was an asymmetry in the use of final lengthening: In both comparisons, final lengthening was predictive only for the language with the greater magnitude of final lengthening. This is in keeping with expectations for English-speaking listeners, who habitually hear segments lengthened substantially in phrase/utterance-final position (e.g., Wightman et al., 1992) and use these effects for interpretation of prosodic structure (e.g., Price et al., 1991).

In a preliminary study (White, Mattys, Series, & Gage, 2007), in which we used an AAX discrimination paradigm with rate-controlled *flat sasasa* utterances, we found that English listeners were able to distinguish both EngS vs EngW and EngS vs Sp, but, unlike the current study, were unable to distinguish EngO from EngW. Critically, however, the utterances in that study were truncated down to ten syllables each, with the final stressed syllable and subsequent segments removed. Thus, the White et al. study and the present one converge in showing that utterance-final syllables, preserved or eliminated, can have a decisive impact on language categorization.

Class discrimination hypothesis

The class discrimination hypothesis was that only languages from different rhythm classes should be distinguished. Thus, there should be particular characteristics of "stress-timed" languages that consistently serve to distinguish them from "syllable-timed" languages, but not from each other. However, we found that English listeners could distinguish different varieties of English, EngS vs EngW, and EngO vs EngW, on the basis of durational cues alone. We must therefore reject the class discrimination hypothesis.

It could be argued that rhythm scores indicate that EngW is in fact "syllable-timed," which could account for listeners' ability to distinguish it from EngS without undermining the class discrimination hypothesis. However, apart from the circular nature of such an argument, and the fact that EngW was almost equidistant in terms of rhythm scores from EngS and Spanish (Fig. 1), this interpretation would fail to account for listeners' discrimination of EngO from EngW. Furthermore, White et al. (2007) found Spanish to be discriminated from EngW. In sum, our results cannot all be explained in terms of a consistent rhythm class account, unless multiple additional classes are proposed.

As discussed above, previous studies have been taken to provide support for rhythm classes on the basis of discrimination between but not within rhythm classes. We argue that absence of discrimination within classes results from the fact that the timing differences that allowed EngW to be distinguished from EngS and from EngO in the present study were absent or attenuated in those studies. To consider, for example, Dutch and English, found by Ramus et al. (2003) not to be distinguishable by French listeners on the basis of *flat sasasa* utterances. As shown in Fig. 1, Standard Dutch and Standard Southern British English have similar contrastive rhythm scores. Furthermore, they both manifest substantial final lengthening effects (e.g., Cambier-Langeveld, 2000). Intrinsic differences in rate are also likely to be smaller between Dutch and English, given their similar phonotactics, than between, for example, English and Spanish (e.g., Dauer, 1983). Thus, the timing differences between languages that listeners have been shown to use in this study – rate, interval contrast, final lengthening - were very likely absent in the Dutch vs English comparison.

Taken together, the data reported here present a strong counter-argument to the class discrimination hypothesis: We have shown that listeners can distinguish not just between different languages of a particular rhythm class, but between accents of a single language, English, on the basis of timing information alone. It should be noted, however, that the participants in the present study were native English speakers, which raises the question of whether categorization is affected, and potentially facilitated, when varieties of the listener's native language are included in

the comparison. There is some developmental evidence, using intact natural speech stimuli, to support that view. For example, Nazzi, Jusczyk, and Johnson (2000) found that 5-month-old American English-learning infants could discriminate between Dutch and British English and between British and American English, but they could not discriminate other languages from the same rhythm class (Dutch vs German) or a different rhythm class (Italian vs Spanish). However, that study used intact speech, whereas our participants, who heard *sasasa* stimuli, had no way of knowing whether they were listening to their native language or to a completely unfamiliar language. Indeed, when asked for their reflections on the task after the experiments, none of them reported any awareness of the linguistic origins of the *sasasa* speech. Furthermore, they exploited the same types of timing distinctions (vocalic and consonantal interval variation, final lengthening) to distinguish English accents and to distinguish between English and Spanish. This indicates that within-native-language categorization is not based on distinct perceptual cues and so is unlikely to represent a special case.

Given that, for rate-normalized stimuli, categorization performance was better between rhythm classes than within, a weaker version of the concept of rhythm class could be retained. If rhythm class simply relates to constellations of features such as syllable phonotactics and the durational marking of stress, there is scope for these to differ more between some languages than others. However, the notion of class is intended to be a categorical one, and so there must be some independent means of establishing where the boundaries lie.

As observed above, durationally-based metrics of rhythmic contrast indicate gradient variation between languages. Furthermore Loukina, Kochanski, Rosner, Shih, and Keane (2011) found that different combinations of rhythmic metrics give rise to different language groupings. Thus, considering the acoustic and perceptual evidence together, the categorical concept of rhythm class has little support.

In this regard, it may be noted that the functions of contrastive rhythm differ greatly between languages that have been traditionally regarded as cohabiting a rhythm class. Consider, for example, two canonical syllable-timed languages: Native Spanish listeners must use stress placement as a cue to lexical identity – there are numerous minimal pairs distinguished only by stress – whilst French lacks stress at the lexical level and indeed native French speakers have been proposed to be "deaf" to stress contrast altogether (Dupoux, Pallier, Sebastián-Gallés, & Mehler, 1997).

Implications for language processing and language development

The reason that listeners, when presented with unfamiliar *sasasa* stimuli, naturally exploit cues such as rate, durational contrast, and final lengthening for language categorization is probably that those same cues are also important for everyday language processing, in particular, for the segmentation of speech into words and phrases (e.g., Christophe, Gout, Peperkamp, & Morgan, 2003; Price et al., 1991). Indeed, boundaries in speech are associated with distinct lengthening effects, with word-initial lengthening localized on the syllable onset, and phrase-final lengthening on the rhyme (e.g., White, 2002). Thus, for the purposes of speech segmentation, English listeners habitually exploit durational information from these two sub-syllabic constituents separately. In the present experiments, this division was reflected in the separate predictive power of metrics of vowel and consonant variation. Variation in the duration of whole syllables was not exploited probably because the timing effects that listeners normally experience and interpret are not evenly distributed over whole syllables.

Sensitivity to overall speech rate also has a role in segmentation, particularly for the interpretation of lengthening effects as prosodic boundary cues. To recognize variation in the duration of subsyllabic constituents as linguistically meaningful, the listener must have a prior expectation about normative segment durations. Attendance to ongoing speech rate is a means by which such expectations can be generated. Indeed, Reinisch, Jesse, and McQueen (2011) demonstrated the importance, for the segmentation of ambiguous stimuli, of listeners' judgement of speech rate based on preceding context.

Like adults, neonates can distinguish languages on the basis of speech in which segmental information is largely eliminated (e.g., Nazzi et al., 1998). On the assumption that adults and infants are sensitive to the same set of timing cues, a testable hypothesis is that infants exposed to two languages within the same rhythm class should be able to distinguish them if the languages differ in speech rate, in durational contrast between vowels and between consonants, and/or in final lengthening.

If such discrimination ability were demonstrated in neonates or young infants, it would indicate perceptual sensitivity to timing cues that could also be exploited for speech segmentation. Boundaries between words and between phrases can indeed be detected in the first year of life (e.g., Christophe, Dupoux, Bertoncini, & Mehler, 1994; Gout, Christophe, & Morgan, 2004), but the precise perceptual abilities exploited for segmentation remain uncertain.

Finally, while durational variation is a key component of speech rhythm, everyday speech contains additional prosodic cues to distinguish languages, and further research should consider the perceptual exploitation of such cues in parallel. Ramus and Mehler (1999) found that *sasasa* speech with natural intonation conferred no advantage over *flat sasasa* in English vs Japanese discrimination by French listeners, but intonation may provide more information for other language comparisons or other listener groups. Likewise, given that loudness has been shown to be fundamental, with duration, to the perception of prominence in English, (Kochanski, Grabe, Coleman, & Rosner, 2005), cross-linguistic variation in loudness contrast may provide an additional cue to language distinctions. However, in support of the importance of durational cues, English listeners did not discriminate two accents of Italian that differed in speech rate any better when presented with natural utterances than with *flat sasasa* utterances (White et al., 2012).

Future studies should also examine how native speakers of other languages exploit the cues we have shown to be

important to English speakers in our study. As durational variation is used very widely as a cue to stress, and perhaps universally as a cue to word and phrase boundaries, we expect that the same set of cues will also be used for categorization by non-English listeners. However, native language experience may *modulate* listeners' sensitivity to durational variation. For example, as discussed above, Spanish has prosodic timing effects, such as final lengthening, that are much smaller in magnitude than those of English, and so Spanish listeners may acquire, through early language experience, sensitivity to smaller durational variations than those perceived by English listeners.

In summary, the experiments reported here show that listeners can distinguish between languages on the basis of temporal cues alone, not only between "rhythm classes," as found in previous studies, but also within a single language. Thus, we conclude that listeners do not have a categorical perceptual sensitivity to rhythm class, in whatever acoustic form that concept might be instantiated. Rather, we propose that listeners systematically exploit a range of timing cues to language differences: speech rate, durational variation between consonantal intervals and between vocalic intervals, and utterance-final lengthening.

Acknowledgments

This study was made possible thanks to a grant from the Leverhulme Trust (F/00 182/BG) and a Research Training Network grant from the Marie Curie foundation (MRTN-CT-2006-035561). We thank Klaske von Leyden, Juan Manuel Toro, and Rod Walters for help with recordings, and Lucy Series and Suzi Gage for assistance in preparation of experimental materials. We also thank Katalin Mády and Kari Suomi for collaboration on the recordings of Hungarian and Finnish respectively, and for permission to reproduce unpublished data from those studies here. We are grateful to three anonymous reviewers for their very helpful comments on an earlier version of this manuscript.

Appendix A. 1. Sentences on which the *sasasa* stimuli were based

A.1. English

- The supermarket chain shut down because of poor management.
- Much more money must be donated to make this department succeed.
- In this famous coffee shop, they serve the best doughnuts in town.
- The chairman decided to pave over the shopping centre garden.
- The standards committee met this afternoon in an open meeting.

A.2. Spanish

- A mí no me gustaba su coche pequeño y viejo.
- Vicente y Susana van de vacaciones este mes a Escocia.
- A pocos pasos de mi casa está una tienda bonita.
- Un chico me dijo hace poco que no había pasado nada.
- Pienso que todo va bien con mis tíos estas Navidades.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bosch, L., & Sebastián-Galles, N. (1997). Native language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition, 65*, 33–69.

Cambier-Langeveld, T. (2000). *Temporal marking of accents and boundaries*. PhD dissertation, University of Amsterdam.

Cho, M. (2004). Rhythm typology of Korean speech. *Cognitive Processing, 5*, 249–253.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America, 95*, 1570–1580.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. L. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics, 31*, 585–598.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science, 1*, 215–219.

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

Dehaene-Lambertz, G., & Houston, D. (1998). Faster orientation latency toward native language in two-month old infants. *Language and Speech, 41*, 21–43.

Dellwo, V. (2008). The role of speech rate in perceiving speech rhythm. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of fourth conference on speech prosody* (pp. 375–378), Campinas.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A destressing 'deafness' in French? *Journal of Memory and Language, 36*, 406–421.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vreken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the international conference on spoken language processing* (pp. 1393–1396), Philadelphia.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 521–602). Oxford: Wiley-Blackwell.

Frota, S., & Vigário, M. (2001). On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. *Probus, 13*, 247–275.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language, 51*, 547–567.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence. Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*, 1038–1054.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., et al. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research, 52*, 1334–1352.

Loukina, A., Kochanski, G., Rosner, B., Shih, C., & Keane, E. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America, 129*, 3258–3270.

Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English. *Language and Speech, 43*, 377–401.

Mehler, J., & Christophe, A. (1995). Maturation and learning of language during the first year of life. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 943–954). Cambridge, MA: Bradford Books/MIT.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143–178.

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

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

Nespor, M. (1990). On the rhythm parameter in phonology. In I. M. Roca (Ed.), *Logical issues in language acquisition* (pp. 157–175). Dordrecht: Foris.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*, 2956–2970.

Prieto, P., Vanrell, M. M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. *Evidence from Catalan, English, and Spanish*. Speech Communication.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics, 35*, 353–362.

Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15th international congress of phonetic sciences, Barcelona* (pp. 337–342).

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

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

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 978–996.

White, L. (2002). English speech timing: A domain and locus approach. PhD dissertation, University of Edinburgh. <http://www.cstr.ed.ac.uk/projects/eustace/dissertation.html>.

White, L. (2012). Where is the rhythm in speech? Paper presented at Making Sense of Sound 2012, Plymouth University.

White, L., Mattys, S. L., Series, L., & Gage, S. (2007). Rhythm metrics predict rhythmic discrimination. In J. Trouvain, & W. J. Barry (Eds.), *Proceedings of the international congress of the phonetic sciences* (pp. 1009–1012).

White, L., & Mattys, S. L. (2007a). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35*, 501–522.

White, L., & Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in romance phonology. Current issues in linguistic theory series* (pp. 237–257). Amsterdam: John Benjamins.

White, L., Payne, E., & Mattys, S. L. (2009). Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In M. Vigario, S. Frota, & M. J. Freitas (Eds.), *Phonetics and phonology: Interactions and interrelations* (pp. 137–158). Amsterdam: John Benjamins.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*, 1707–1717.