Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies

Tamara V. Rathcke*

English Language and Linguistics

University of Kent

Cornwallis North West

Canterbury CT2 7NF

United Kingdom

Rachel H. Smith

English Language

University of Glasgow

12 University Gardens

Glasgow G12 8QH

United Kingdom

* T.V.Rathcke@kent.ac.uk

Running title: Acoustic bases of rhythm typologies

ABSTRACT

Research into linguistic rhythm has been dominated by the idea that languages can be classified according to rhythmic templates, amenable to assessment by acoustic measures of vowel and consonant durations. This study tested predictions of two proposals explaining the bases of rhythmic typologies: the *Rhythm Class Hypothesis* which assumes that the templates arise from an extensive vs. a limited use of durational contrasts, and the *Control and Compensation Hypothesis* which proposes that the templates are rooted in more vs. less flexible speech production strategies. Temporal properties of segments, syllables and rhythmic feet were examined in two accents of British English, a 'stress-timed' variety from Leeds, and a 'syllable-timed' variety spoken by Panjabi-English bilinguals from Bradford. Rhythm metrics were calculated. A perception study confirmed that the speakers of the two varieties differed in their perceived rhythm. The results revealed that both typologies were informative in that to a certain degree, they predicted temporal patterns of the two varieties. None of the metrics tested was capable of adequately reflecting the temporal complexity found in the durational data. These findings contribute to the critical evaluation of the explanatory adequacy of rhythm metrics. Acoustic bases and limitations of the traditional rhythmic typologies are discussed.

4370Fq Acoustical correlates of phonetic segments and suprasegmental properties

I. INTRODUCTION

A The rhythm class hypothesis

The generally recognized and widespread typology of linguistic rhythm is fairly simple. It dates back to Lloyd James (1940) who noted that the perceptual impression of rhythm in languages like Spanish and English could be described either as 'machine-gun' or as 'morse-code', respectively. This impressionistic notion of linguistic rhythm was subsequently elaborated into a typological theory of two linguistic rhythm classes (Abercrombie, 1967; Pike, 1945): two groups of languages were identified on the basis of assumed perceptual templates, a 'syllable-timed' group with a 'machine-gun rhythm' (e.g. Spanish, Italian or French) and a 'stress-timed' group with a 'morse-code rhythm' (e.g. English, Dutch or Russian).[1] The theoretical simplicity and transparency of the rhythmic typology has, however, often been obscured by the practical issue of defining the acoustic features responsible for the maintenance of the two rhythm templates, and even the typological issue of deciding which languages belong with which template.

Some early approaches assumed that the perceptual templates were derived from two different sources of isochronous regularity in the acoustic signal: evenly spaced syllables in syllable-timed languages vs. equal inter-stress intervals in stress-timed languages (Abercrombie, 1967; Pike, 1945). However, any attempt to find evidence for isochronous duration patterns in speech has failed (e.g. Dauer, 1983; Roach, 1982). Only under very severe linguistic constraints on the form and function of sentences (similar syntax, controlled syllabic and segmental composition of words), do regular, although not strictly isochronous, units appear on the acoustic surface, shaped mainly by prosodic factors, i.e. rhythmic foot structures and their phrasal positions (Lehiste, 1977; Classe, 1939). In consequence isochrony has been abandoned or considered a perceptual construct arising from listeners' ability to compensate for predictable acoustic regularities such as successive lengthening of rhythmic feet towards the end of prosodic phrases (Lehiste, 1977).

When rejecting the isochrony hypothesis, Dauer (1983) noticed that languages traditionally called syllable-timed displayed little or no vowel reduction and restricted consonantal phonotactics, in contrast to those presumed to be stress-timed, which had a high level of vowel reduction (i.e. a predominance of schwas) in unstressed syllables and allowed complex consonant clusters. The observed correlation between proposed rhythm class and properties of segmental phonology and phonotactics gave rise to the development of a number of so-called rhythm metrics, motivated by the possibility of acoustically separating languages from the different rhythm classes based on temporal indices capturing durational variability of vocalic and consonantal intervals (Ramus, Nespor, and Mehler, 1999). Acoustic measures were further shown to support the rhythm class distinction among dialects of a language (Low, Grabe, and Nolan, 2000): here, the amount of vowel reduction in unstressed syllables achieved a particularly good separation between Singapore English (syllable-timed, with more peripheral vowels in unstressed syllables) and British English (stress-timed, with more centralized vowels in unstressed syllables). This approach appeared very promising, as the first acoustic measurements undertaken on a small number of languages and dialects indeed supported their presumed rhythm class affiliation.

The initial success of rhythm metrics has been followed by empirically justified criticism since their ability to provide an acoustic basis for a clear separation between two rhythmic templates has not stood up to the cross-linguistic data. There are three basic groups of rhythm metrics, all aimed at identifying the most reliable measure of rhythm class separation: (1) a vowel-to-consonant ratio, expressed as a percentage (%V, Ramus *et al.*, 1999); (2) the standard deviation Δ, expressed in ms (Ramus *et al.*, 1999), and its normalized counterpart the variability coefficient Varco (Dellwo and Wagner 2003), calculated as the standard deviation divided by the mean duration; (3) the pairwise variability index PVI, calculated as the mean temporal distance between pairs of successive intervals (raw or normalized to the mean duration of those two intervals, Low *et al.*, 2000). The metrics have been applied to vocalic, consonantal, (pseudo-)syllabic and foot intervals (among many, Deterding, 2001; Grabe and Low, 2002; Low *et al.*, 2000; Nolan and Asu, 2009; Ramus *et al.*, 1999; Rathcke and Smith,

2011; White and Mattys, 2007). Although most of these metrics have repeatedly confirmed that, in fact, assumed syllable-timed languages have less variable durational patterns and a greater proportion of vocalic material than assumed stress-timed languages, the metrics' success in validating of the originally proposed rhythmic templates has remained questionable. Just to name a few critical outcomes of some recent studies, a given language might be placed in different classes based on the output of different metrics (Grabe and Low, 2002). The differences in metric scores induced by variation in materials, speaking styles, speech rate or speaker identity can exceed those related to rhythm class affiliation (Arvaniti, 2009; Barry, Andreeva, and Koreman, 2009; Wiget *et al.*, 2010). Different rhythmic affiliations sometimes fail to show up in calculations of common metrics (White, Payne, and Mattys, 2009). And crucially, cross-linguistic comparisons revealed that the relationship between the traditional vocalic metrics (%V and ΔV) and the presence of reduced, centralised vowels in the phonemic repertoire of a language is not as straightforward as originally assumed (Easterday, Timm and Maddieson, 2011). These findings suggest that despite a sustained interest in the rhythm class hypothesis, the factors responsible for the assumed rhythmic templates still have not been reliably identified, which makes the validity of metrics questionable (cf. Knight, 2011).

B The control and compensation hypothesis

An alternative rhythmic typology was proposed by Bertini and Bertinetto (2010). The dichotomy of controlling and compensating languages is related to the original assumption of two rhythmic templates but is framed within a more explanatory approach. The control and compensation hypothesis states that languages traditionally described as syllable- vs. stress-timed employ two opposite speech production strategies (accordingly, the terms 'syllable-timed' and 'stress-timed' were replaced with 'controlling' and 'compensating' in order to prevent potential misinterpretations of the assumptions made by the new model). Controlling languages are assumed to allow for a limited degree of coarticulation, i.e. those languages "control" for the amount of articulatory energy spent on production of each and every segment,

leading to lesser reduction and a smaller degree of overlap between consonant and vowel gestures. In contrast, compensating languages are assumed to display a higher degree of coarticulation and various compensation strategies to adjust e.g. for an increasing number of segments within syllabic units. The typological terminology therefore derives from two types of speech production behaviour, with languages having a strong preference for either the one or the other. However, fully controlling or fully compensating speech production is deemed an abstraction while the target of research interest has been placed on the continuum between the two (hypothetical) poles.[2] The authors acknowledged that the positioning of a language along the continuum may partly be influenced by its phonotactics: languages with simpler phonotactic structures are more likely to adopt the controlling strategy while languages with rich phonotactics require a more flexible, compensating articulatory setting.

These ideas were developed with reference to articulatory gestural coordination models (esp. Fowler, 1977; *et seq.*). The empirical evaluation of the hypothesis, however, has not involved articulatory investigation but has followed the way paved by the rhythm metrics approach: it involves measuring acoustic durations, using the control and compensation index (CCI) specifically designed to address the hypothesis. The CCI is based on the principle of the PVI. Both indices operate locally and capture the degree of durational contrast realized between two adjacent intervals of the same type. The innovation of the CCI is to take into account the number of segments composing the measured intervals, i.e. the duration of each interval is divided by the number of its components before the pairwise variability is calculated. Thus the index gives information about the variability associated with the average amount of time spent on the production of each segmental unit (consonant or vowel) in running speech. Compensating languages are expected to show higher variability than controlling languages overall, and in particular, to show higher variability for vocalic than for consonantal intervals due to a high level of vowel reduction. In contrast, controlling languages are predicted to show a comparable level of variability on both intervals and to reduce the duration of all segments in a more proportional way under variations of speech rate.

The hypothesis underlying the CCI is strongly connected to the well-known phenomenon of constituent-internal compression, or compensatory shortening (Fowler, 1981). This type of compression describes a negative correlation between the duration of larger structural units (syllables or feet) and their components (segments or syllables), i.e. when the duration of a constituent increases due to its containing a higher number of subunits, the average duration of a subunit tends to decrease. Compression effects at an intra-syllabic level are relatively well studied. There is some cross-linguistic evidence that vowels are shortened in syllables with more consonant segments, regardless of whether the consonants occur in onset or coda position (e.g. Katz, 2012; and references therein). As for the consonants themselves, their position in the syllable is crucial (Byrd, 1995): consonants usually exhibit a stronger compression effect in onset than in coda clusters, which is usually explained in terms of in-phase (C-V) vs. anti-phase (V-C) coupling of consonantal and vocalic gestures (Browman and Goldstein, 1989). Also, the relative phase and the resulting overlap of gestures are more stable in onset than in coda clusters (Byrd, 1996). This important aspect of consonantal timing is however not taken into account in the CCI proposal, as the durational variability of an average consonant is calculated on the basis of a cluster interval which is not further distinguished as consisting of onsets, codas or both. It seems that in contrast to vocalic compression, the articulatory underpinning of consonantal variability is generally less clearly elaborated in relation to the proposed dichotomy between control and compensation.

C Rhythm typologies and linguistic rhythm

The core difference between the two typological perspectives lies in their different assumptions about the origin of the two rhythmic templates: they are suggested to arise either (1) from an extensive use of duration as a contrastive cue for various linguistic functions like stress and vowel quality (the *rhythm class hypothesis*, RCH) or (2) from articulatory constraints on the flexibility of coarticulation patterns (the *control and compensation hypothesis*, CCH). However, it is not quite clear whether these two accounts are mutually exclusive or describe two co-existing strategies whose combined effects give rise

to a unified rhythmic typology. The latter appears to be a valid assumption since RCH and CCH do not differ in terms of the groupings of languages in two classes that they propose. Before attempting to outline testable predictions congruent with each of the theoretical proposals (see below, Sect. I.D), we discuss a number of limitations in the way rhythm is dealt with according to both hypotheses.

First, the focus on cross-linguistic comparisons has tended to obscure an important issue, namely the fact that timing alternations in the acoustic signal can be driven by two sources, the underlying phonological system and its phonetic implementation. For example, vowel reduction as well as quantity contrasts can be expressed by duration and/or by quality – two phonetic exponents of phonological systems which are usually assumed to be equivalent in traditional RCH-inspired research. This might explain the otherwise confusing finding that contrastive vowel length and especially phonological vowel reduction to schwa were not straightforwardly reflected in %V and ΔV metrics when applied to 22 languages with systematic differences in their vowel phonologies (Easterday *et al.*, 2011). To evaluate the contribution of phonetic implementation separately from that of the phonological system, we need to compare two varieties of a language, which have the same basic phonology but different rhythm class affiliations. Cross-dialectal comparisons likewise have the potential to inform development of the control and compensation hypothesis by allowing the contribution of articulatory strategies to the rhythmic typology to be estimated. A strong argument in favour of the articulatory-based perspective would be a finding that a controlling (syllable-timed) variety tends to simplify consonant clusters by dropping segments or inserting vowels.

Second, a major question which has remained largely untouched is how exactly the low-level timing effects emphasized in both proposals are related to rhythm in speech and language. A broad definition of rhythm, free of a pre-set connection to a regular pattern resulting in isochrony, focuses on temporal alternations between strong and weak elements as well as their groupings into larger units (cf. Patel, 2008). That is, linguistic rhythm and rhythmic typology may be related to the structure and implementation of the prosodic hierarchy (e.g. Fletcher, 2010). Obviously, to address this aspect of

linguistic rhythm we need to look beyond durational variability at the segmental level, yet neither typological view focuses on feet as potentially important rhythmic units in a syllable-timed/controlling language. This issue was raised by Nolan and Asu (2009) who applied the PVI-concept to syllables and rhythmic feet in several languages and argued on the basis of their findings that the two rhythmic patterns, syllable- and stress-timing, could coexist in a language and operate at different levels of the prosodic hierarchy. Similarly, the higher structural level of the rhythmic foot has proved extremely relevant for understanding constituent-internal compression which occurs exclusively in stressed syllables, shortened by following unstressed syllables (Fowler, 1981; White and Turk, 2010). An important finding is moreover that languages of the two traditional rhythm classes differ in this respect: e.g. Vayra, Fowler, and Avesani (1987) observed compensatory shortening of strong syllables in English but not Italian feet of various sizes. Moreover, rhythmically divergent varieties of the same language have similarly been shown to differ in the degree of foot-level compression: with each additional syllable, the foot duration in syllable-timed Indian English speech increases more than in stress-timed American English (Krivokapić, 2013). Rather surprisingly, the CCH does not make an explicit attempt at distinguishing between intra-syllabic and intra-foot levels of compression (although the CCI-metric should, if applied to syllabic or foot intervals, pick up on compression at either level).

Recently, the underpinnings of the two putative rhythm classes have begun to be explored in the context of prosodic timing (e.g. Prieto *et al.*, 2012; White *et al.*, 2009): languages or dialects of the stress-timed class have been shown to implement stronger temporal demarcation of accented and phrase-final syllables in comparison to those from the syllable-timed class which make restricted use of duration for the purposes of high-level prosodic functions. Preliminary evidence of this kind suggests that there is an interdependence of rhythm class affiliation and language- or dialect-specific prosodic timing, which needs to be integrated into the traditional RCH-view.

D Aims and predictions of this study

The general aim of the study was to explore the acoustic bases of the rhythmic typologies deriving from the *rhythm class* and *control and compensation* hypotheses, using a carefully controlled dataset that kept linguistic materials constant, while allowing phonetic implementation to vary. More specifically, we wished to test the core ideas of the two hypotheses, i.e. maximization of durational contrasts vs. constituent-internal compression, and to establish whether there is evidence for one, both, or neither proposal. Further, we wished to understand the relationship between rhythm metrics proposed in accordance with the two typologies and linguistic factors which are known to affect speech timing, and to calibrate the metrics' ability to reflect a subtle rhythmic difference between two closely related varieties of a language.

We chose two varieties of Yorkshire British English as spoken by monolinguals from Leeds and by Panjabi-English bilinguals from Bradford. Leeds and Bradford are cities 9 miles apart, sometimes considered part of a single large urban zone within West Yorkshire. With 26% of its population of Asian ethnic identity (as compared to 8% in Leeds) Bradford is home to half of the South Asian population in Yorkshire, and has one of the largest South Asian communities in the UK. Importantly, bilingual speakers from Bradford have been suggested to sound relatively syllable-timed despite the fact that Panjabi itself has never been classified as a syllable-timed language (Heselwood and McChrystal, 2000). Syllable-timing has moreover been often demonstrated to constitute the rhythmic basis in contact varieties of English (Deterding, 2001; Torgersen and Szakay, 2011; Low *et al.*, 2000). The Bradford variety, not yet analysed within the RCH paradigm, seemed therefore to be a perfect candidate to study the acoustic origins of a subjectively perceived rhythm class affiliation.

We formulated two sets of predictions in the spirit of the two rhythm typologies. (1) Following and extending the ideas of the RCH, we would expect Bradford to display a fairly weak temporal reflex of vowel phonology and a lesser degree of accentual and phrase-final lengthening. The finding of fewer

cases of syllables with reduced vowel qualities in metrically weak syllables in Bradford would further support the RCH view. In contrast, the Leeds variety was hypothesized to show a stronger temporal cueing of the vowel system, prominence and prosodic hierarchy, and to have a higher proportion of weak syllables with reduced vowel qualities. (2) In keeping with the CCH, we expected to observe limited or even missing compression effects in Bradford. On the contrary, Leeds speech was expected to display considerable compression effects on vowels as well as consonants in complex syllables, and on prominent syllables beginning longer rhythmic units such as rhythmic feet. Also, the finding of a smaller number of syllables with complex consonant clusters and/or a smaller number of long rhythmic feet in Bradford compared to Leeds would support the articulatory perspective suggested by the CCI. As outlined in I.C above, there was no obvious reason to preclude the possibility that articulatory forces and linguistic functions could simultaneously contribute to the acoustic bases of the rhythmic templates. The issue was considered a matter for empirical investigation, as addressed below.

We will also compare the output of the rhythm metrics proposed by the two accounts with respect to how sensitively they can differentiate between the two varieties and pick up on the phonetic substantiation of the linguistic functions and articulatory forces listed above. In sum, each section of the results below (consonants in III.A, vowels in III.B and higher-level structures in III.C) will report data regarding the following three measurements: (1) rhythm metrics, (2) distributions and (3) durations.

II. METHOD

A Participants and recordings

The database consisted of nursery rhymes ("Jack and Jill", "Humpty Dumpty" and "Baa Baa Black Sheep") and "The Princess and the Pea" passage, a well-known fairy tale. The rhymes had a fairly simple phonotactic and phonological structure and constituted the most rhythmical materials in the study, making them especially suitable to test for the effects of prosodic structure and foot-level compression. The passage featured more complex phonotactics and analysis of it was expected to be particularly useful

for estimating the role of phonetic realizations of vowel categories and consonant clusters, and for testing the effects of syllable-level compression. Taken together, these materials were deemed appropriate to test the hypotheses put forward in I-D. Note that the size of our corpus far exceeded the sentence lists commonly used as corpora to calculate rhythm metrics.

We recorded speech from six middle class speakers (three males, three females) for each variety. The recordings were made at the universities of Leeds and Bradford during summer-autumn 2010. The Bradford participants were aged between 31 and 41 (mean age: 36). They were born and grew up in Bradford (one male speaker was born in Pakistan and moved to Bradford as a toddler). All participants were embedded in a bilingual environment, speaking Panjabi with their parents (who were originally from Pakistan) and English with peers, siblings and at work. Most spoke a third language (Urdu or Arabic). The Leeds speakers were aged between 26 and 34 (mean age: 30). Most participants had lived in their home city all their lives, and all of them confirmed that they spoke with the local accent.

To follow up on previous impressionistic descriptions of the rhythmic affiliation of the two varieties of British English (Heselwood and McChrystal, 2000), we ran a small-scale perception study. Ten native English-speaking phoneticians (with expert knowledge of the RCH), based in the UK, were asked to rate short samples of each speaker's productions on a 120 mm long continuum spanning the two poles, *"strongly syllable-timed"* on the left and *"strongly stress-timed"* on the right. The expert listeners were asked to listen to each sample as often as necessary, and to place a star on the continuum to express their perceptual impression. The samples for this test were taken from participants' readings of two passages ("The Princess and The Pea" and "The Sailor", the latter not analysed acoustically in this paper). Six short extracts from each passage were used. Each extract was represented by a sample from one Bradford and one Leeds speaker (6 seconds on average, ranging from minimally 4 to maximally 8 seconds per speaker). The 24 target samples (12 speakers x 2 extracts) were supplemented by 12 filler samples which consisted of similar-sized extracts from the "Cinderella" passage recorded in the 1990s for the *Intonational Variation in English* (IViE) corpus (Grabe, 2001). Readings by one male and one

female speaker from Bradford (Panjabi-English bilinguals), Belfast, Cambridge, Leeds, London (speakers of West Indian descent) and Newcastle were included as fillers. Three randomization lists for the total of 36 test stimuli were used. Expert listeners' responses, measured in mm, were first transformed to a z-score, and subsequently subjected to linear mixed effects modelling with *dialect* of the speaker as the predictor and *listener*, *speaker*, *extract* and *order* of presentation as the random intercepts.

The results confirmed that the bilingual speakers from Bradford who participated in our study sounded more syllable-timed than our monolingual speakers from Leeds ($\chi^2(1) = 24.2$, $p < 0.05$). The difference amounted to 0.6 z-scores (with Bradford speakers' extracts being placed on average 0.2 standard deviations below the listener's individual mean and Leeds speakers' extracts being positioned on average 0.4 standard deviations above the listener's mean). In terms of distance along the continuum, samples produced by Bradford speakers were located around 62% while Leeds speakers' extracts were placed around 72% of the scale (where 0% = strongly syllable-timed and 100% = strongly stress-timed). Despite a general bias towards the stress-timed end of the continuum, there was a slight yet clear perceptual difference between the speakers from the two groups in terms of their rhythm class affiliation.

B Segmentation and labeling

The database was created using *EMU Speech Database* (Harrington, 2010). Segmentation and labeling decisions were based on acoustic data (waveform, spectrograms, and f0 trajectories) as well as auditory impression. We only analysed fluently spoken stretches, i.e. we excluded phrases containing false starts, repairs, mispronunciations or hesitations. We segmented vocalic and consonantal intervals as well as syllables. Each interval was labelled with the number of segments it contained. Consonantal intervals were additionally marked as belonging to syllabic onsets or codas. Vowels were classified auditorily (based on their realisation rather than phonological status) as diphthongs, tense, lax or reduced monophthongs. Voiceless vowels and sonorants were not included in vocalic intervals, but were treated as constituting syllable nuclei where appropriate. The segmentation of syllables was guided by the

maximal onset principle, which posits that consonant sequences are assigned to syllable onsets as long as they do not form phonotactically illegal clusters (Selkirk, 1981). This principle was applied within prosodic phrases irrespective of morphological or lexical boundaries. The guidelines discussed in Peterson and Lehiste (1960) and refined by White and colleagues (White and Mattys, 2007, White *et al.*, 2009) were used for segmentation of vocalic and consonantal intervals to ensure comparability across the studies. Postpausal initial stops were allocated a notional 40 ms closure duration to prevent syllables from appearing consistently shorter phrase-initially than in other positions. Glottal stops and cases of strong glottalisation were treated as individual segments whereas weak glottalisation and weak nasalisation (i.e. in cases where weak or no spectrographic cues supported the auditory impression) were considered as a secondary articulation of the vowel and not marked separately. Consonant insertions, identified on auditory and acoustic grounds, were counted as additional segments. In cases of a complete place and manner assimilation of two adjacent segments, only one segment was labelled.

For each syllable, we identified its metrical stress, phonotactic composition, number of segments, phrasal position and prominence. Prominence comprised two levels (accented/unaccented). The metrical status of each syllable (strong/weak) was assigned based on the principles of metrical stress theory (Hayes, 1995). Phonotactic composition was specified in terms of consonant and vowel constituents as CV, VC, CCVC, etc (this was needed for the correct classification of consonantal intervals as constituting either onset or coda). Phrasal position was specified as medial, initial or final (the latter two being limited to one syllable only). Foot intervals were defined as beginning with a stressed or accented syllable and continuing until the end of the phrase or until the next stressed or accented syllable (cf. Abercrombie, 1967).

Basic segmentation and labelling were carried out by a junior phonetician and checked for consistency by the first author, who also added information about prosodic structure (prominence and phrasing). The second author checked the prosodic labelling and identified vowel quality.

C Metrics

Since accents of British English have been shown to differ mostly on rate-normalised vocalic metrics such as %V, VarcoV and (to a lesser extent) nPVI-V (White and Mattys, 2007), we calculated these three metrics for vowel intervals. Despite a strong correlation with speech rate, non-normalised metrics such as ΔC and rPVI-C have been suggested to reflect differences related to rhythm class affiliation better than their normalised counterparts VarcoC and nPVI-C, which "eliminate almost all the critical linguistic variability" (White and Mattys, 2007:518). We therefore included both the non-normalised ΔC, rPVI-C and the normalised metrics VarcoC, nPVI-C for consonants. Few studies have applied metrics to rhythmically high-level structures such as syllables and feet (cf. Nolan and Asu, 2009; Rathcke and Smith, 2011), so we decided to concentrate on rate-normalised Varco and PVI metrics for these structures. Additionally, a CCI-index was calculated for all durational intervals investigated (vowels, consonants, syllables and feet). The following CCI-equation was used:

(1) $$CCI = \sum_{k=1}^{m-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right| / (m-1)$$

Here and below, *n* refers to the number of subunits in the interval, *m* to the total number of intervals; *d* corresponds to the interval duration; *k* represents the serial order of the interval. (1) derives from the raw PVI-formula in (2) below and introduces an additional parameter, the number of segments (or more generally speaking, subunits) per interval. The underlying principle of both metrics is similar, since they calculate a mean durational difference between successive intervals of the same type.

(2) $$rPVI = \sum_{k=1}^{m-1} \left| d_k - d_{k+1} \right| / (m-1)$$

The rate-normalised version of the PVI-metric is derived by dividing the difference between successive intervals by their mean duration, as shown in (3).

(3) $$nPVI = \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1)$$

The vowel-to-consonant ratio %V was calculated by dividing the total duration of vocalic intervals by the total duration of vocalic and consonantal intervals and multiplying the result by 100. Varco is the standard deviation of interval durations divided by their mean, while ΔC is simply the standard deviation of consonantal intervals.

D Analyses

We used by-subject analyses of variance to test for dialectal influences on rhythm metrics, and also on the distribution of vocalic, consonantal, syllabic and foot-level features, e.g. vowel reduction in weak syllables (percentage of vowels realized as full vs. reduced), presence of syllabic consonants (percentage of syllabic nuclei realized as sonorant consonants), complexity of consonant clusters (percentage of consonant intervals with less vs. more than three consonants), syllable/foot complexity (percentage of syllables/feet realized with more than three subunits). To ensure that these results do not merely reflect speech rate differences between speakers of the two dialects, we calculated speaker-specific rates of speech (in syllables per second) and compared them using a by-subject analysis of variance. The analysis showed very similar speech rates across the speaker groups (mean rate of 5.6 syll/sec in Leeds and 5.4 syll/sec in Bradford with standard deviations of 0.5 in each case) and no significant effect of dialect ($F(1,10)=0.57$).

To understand the durational patterns, linear mixed effects models were fitted (by maximum likelihood). Modelling via a backward fitting procedure was performed in R (version 3.1.0) using lme4 and lmerTest packages (i.e. all non-significant predictors were removed one-by-one until the model contained only those predictors which significantly contributed to the model fit). We tested for all possible three-way interactions of the main predictors (five in each model) and established their contributions to the model fit through likelihood-ratio tests. For each significant interaction, the contrasts of interest were treatment-coded, while levels of the factors irrelevant to the interaction were sum-coded. This procedure enabled us to unravel the bases of the interactions.

The predictors tested were partly specific to each dependent variable. The group factor *dialect* (Leeds, Bradford) was included in every model. For each dependent variable, the predictive power of the two prosodic factors *phrasal position* (initial, medial, final) and *accentuation* (accented, unaccented) was tested. *Number of subunits* (segments or syllables) was included as a covariate. In case of vocalic, syllabic and foot intervals, we also tested the effect of *vowel quality* (tense/diphthongal, lax, reduced) on interval duration. For consonants, *structural position* of the interval within the syllable (onset, coda) was additionally tested. Speaker-specific idiosyncrasies were taken into account by defining several speaker terms in the models: in addition to the intercept of *speaker*, a random speaker-specific slope was fitted to each factor entering the model. By including the random slopes, we ascertained that the dialectal differences reported were not driven solely by a subset of the speakers tested. Additionally, the vocalic and syllabic models included the phonotactic composition of a syllable (a factor with 12 levels) as a random effect, nested within the number of segments.

All durations below refer to the estimates from the best-fit models, rounded to the nearest 5 ms. We report the results of best-fit models and concentrate mainly on the differences between dialects; the result patterns that generalize across the two dialects are only elaborated upon where necessary for in-depth discussion of the hypotheses put forward in I.D. To check that the result patterns are not driven by the nursery rhymes only, we re-ran all statistical analyses for consonantal, vocalic, syllabic and foot-level intervals excluding the nursery rhymes, and found that the best-fit models yielded the same set of statistically significant predictors as far as dialectal differences are concerned, thus confirming that the result patterns were not driven by a small part of the corpus which was more rhythmic than the rest.

III. RESULTS

A Vowel patterns

Metrics. Table 1 summarizes the output of four metrics applied to the vowel data. Again, Leeds showed higher variability of vocalic intervals as measured by Varco and PVI, and also a lower percentage of

vocalic parts of the signal (%V). The CCI-V score was lower in Leeds, suggesting a larger amount of vowel compression. However, none of these numerical differences was statistically supported: only the by-subject ANOVA run on %V showed a trend towards significance ($F(1,10)=3.5$, $p=0.09$).

##Table 1 around here##

Distributions. Distributional analyses revealed that Bradford and Leeds speakers did not differ with respect to the frequency with which they realized full vs. reduced vowel categories in metrically weak syllables, with the exception of Leeds speakers producing slightly but not significantly more voiceless vowels as nuclei (2.6% vs. 1.4%).

Durations. The results for vocalic interval durations are given in Table 2. The model fit was improved by one three-way interaction involving *dialect* (*dialect*accentuation*number of segments*, $\chi^2(1)=6.5$, $p=0.01$). The interaction indicated that there were dialect-specific vocalic compression effects which depended on the presence of prominence. Under accentuation, compression of the vowel, i.e. its shortening by 20 ms per consonant added to the syllable, was present in Bradford ($t(21)=2.9$, $p<0.01$) as well as in Leeds ($t(21)=2.7$, $p<0.05$). Significant dialectal differences surfaced in absence of an accent ($t(31)=2.8$, $p<0.01$). In Bradford, compression of the vowel was found exclusively in accented syllables, i.e. there was no indication of vowel compression without prominence ($t(22)=1.1$, n.s.). In contrast, Leeds speakers showed significant vocalic compression effects regardless of accentuation, although a slightly smaller magnitude of compression was found for unaccented vowels (15 ms, $t(21)=2.0$, $p=0.058$).

##Table 2 around here##

B Consonantal properties

Metrics. Five consonantal metrics were calculated for each speaker (Section II.C). The outputs are reported in Table 3. According to all metric results, Leeds demonstrates a greater variability in consonantal intervals. Analyses of variance conducted on the scores revealed, however, that only the Varco-C metric picked up a significant cross-dialectal difference in the realization of consonants

(F(1,10)=7.8, p=0.02). The nPVI-C showed a trend towards significance (F(1,10)=3.7, p=0.08). In contrast, numerical differences in the non-rate-normalised scores ΔC, PVI-C and CCI-C were not supported statistically.

##Table 3 around here##

Distributions. We did not find distributional differences between Bradford and Leeds in terms of the complexity of realized consonant clusters. There were no dialectal differences in the numbers of complex onset intervals (5% for both dialects) or coda intervals (3.5% for both dialects). However, analysis of variance revealed that Bradford speakers produced sonorants as syllabic nuclei significantly less often than Leeds speakers did (0.7% vs. 1.3%, F(1,10)=7.8, p=0.02).

Durations. The results of mixed-effects analyses are presented in Table 4; they revealed two significant effects involving *dialect*. The first significant finding concerned an interaction between *dialect, phrasal position* and *number of segments* in the consonantal interval ($\chi^2(2)=15.7$, $p<0.001$). Speakers of both dialects produced similar, i.e. statistically equivalent, amounts of lengthening per additional consonant in phrase-initial (40 ms, t(496)=6.4, p<0.001) and –medial (50 ms, t(22)=17.0, p<0.001) positions. However, the amount of lengthening differed in phrase-final positions. Whereas Bradford speakers produced 10 ms more lengthening per additional segment to mark finality (phrase-medial/-final comparison: t(9959)=2.3, p<0.05), the durational contrast between phrase-medial and phrase-final clusters was more pronounced in Leeds where we found approximately 25 ms more lengthening per additional consonant in phrase-final positions (t(9956)=7.3, p<0.001). Accordingly, the durational difference between the dialects amounts to 15 ms per consonant in the interval (t(37)=2.7, p<0.01), and increases with a higher number of consonants in phrase-final positions.

Second, *dialect* interacted with *phrasal position* and the *structural position* of consonantal intervals within syllables ($\chi^2(2)=19.7$, $p<0.001$). This interaction indicated that in Bradford, coda intervals in phrase-final positions were 10 ms shorter than in Leeds (t(24)=2.5, p<0.05).

##Table 4 around here ##

These results are interpretable in terms of the RCH. We did not find any indication of a greater amount of compression in consonant clusters produced by Leeds in comparison to Bradford speakers (for either onset or coda clusters): that is, there was no differential shortening of consonant duration due to an interaction between *number of segments* and *dialect*. Rather, we found differential lengthening per consonant in specific phrasal positions, as discussed above.

C Rhythmic structures: syllables and feet

Metrics. The output of rhythm metrics applied to the higher level structures of syllables and rhythmic feet is reported in Table 5. All metrics but one diagnosed a slightly higher variability for Leeds speakers' productions of syllables and feet. Somewhat surprisingly, CCI-F was found to be higher in Bradford than in Leeds. However, none of the observed differences approached significance.

##Table 5 around here##

Distributions. With respect to distributional properties, speakers of the two varieties produced different percentages of syllables with complex segmental composition (i.e., syllables with more than three segments; $F(1,10)=7.0$, $p<0.05$), with Bradford speakers unexpectedly having a slightly higher percentage of complex syllables as compared to Leeds (9% vs. 8.3%). As far as rhythmic feet are concerned, speakers of both varieties produced structures consisting of one to five syllables, with disyllabic (43%) feet being the most frequent foot structure in both dialects, followed by monosyllabic and trisyllabic feet (22-23% each).

Durations. The modelling results for durational differences at the syllabic level are given in Table 6. There were five significant three-way interactions involving *dialect*. First, the interaction of *dialect*, *vowel quality* and *phrasal position* ($\chi^2(4)=31.7$, $p<0.001$) indicated that the dialects differed in the duration of syllables containing tense/diphthongal vowels at phrasal edges. These syllables were 25 ms shorter in phrase-initial positions ($t(31)=2.0$, $p<0.05$) and 25 ms longer in phrase-final positions ($t(24)=2.0$, $p=0.053$) when produced by Leeds in contrast to Bradford speakers. These effects point

towards an enhanced durational marking of the prosodic hierarchy for Leeds long vowels. The same pattern was mirrored in the vowel duration data (tense/diphthongal vowels averaged 105 ms in Bradford and 91 ms in Leeds phrase-initially, but 157 ms in Bradford and 164 ms in Leeds phrase-finally), though the interaction did not reach significance there. The implication is that the durational reflexes of phonological vowel length may not be limited to the vowel itself.

##Table 6 around here##

Second, dialect-specific durations depended on the number of segments per syllable in interaction with three other factors: (1) phrasal position (*dialect*phrasal position*number of segments*, $\chi^2(2)=21.6$, $p<0.001$), (2) the presence of phrasal accent (*dialect*accentuation*number of segments*, $\chi^2(1)=19.5$, $p<0.001$) and (3) the vowel quality (*dialect*vowel quality*number of segments*, $\chi^2(2)=14.2$, $p<0.001$). In phrase-initial and -medial positions, both dialects had similar slopes of 30-40 ms lengthening per segment added to the syllable. Cross-dialectal differences surfaced phrase-finally where Leeds speakers produced 15 ms more lengthening per segment than Bradford speakers ($t(43) = 3.1$, $p<0.01$). That is, we find a stronger prosodic timing effect in Leeds than in Bradford. This effect resonates with the results for consonantal properties discussed above. Somewhat reflecting the results for vowel compression, Bradford speakers produced 10 ms more lengthening per additional segment in unaccented syllables in contrast to Leeds speakers ($t(25)=2.3$, $p<0.05$). Accordingly, in Bradford, the lengthening effect of an increased number of segments was significantly stronger in unaccented syllables than in accented syllables ($t(7323)=7.3$, $p<0.001$), whereas in Leeds, lengthening due to a greater number of segments was similar in all syllables regardless of accentuation ($t(7311)=1.2$, n.s.). This difference can be explained by the lack of vowel compression in unaccented syllables in Bradford, again mirroring the results for vowel compression presented above. And lastly, syllables with reduced vowels received 15 ms more lengthening per additional consonant in Leeds as compared to Bradford ($t(46)=2.6$, $p<0.05$).

The interaction of *dialect*, *accentuation* and *phrasal position* ($\chi^2(2)=20.2$, $p<0.001$) suggested dialect-specific effects for accented vs. unaccented syllables in different phrasal positions. However, planned comparisons did not reveal any significant effects relevant to the hypotheses.

We then analysed durations of the whole rhythmic foot as well as durations of the prominent syllable beginning a foot. There were no significant effects of dialectal variety, either in terms of foot-level compression, or in terms of prosodic lengthening effects (the output of the best-fit model is given in Table 7).

##Table 7 around here##

IV. DISCUSSION

A Cross-dialectal timing differences and metric scores

The present study investigated the relationship between rhythm templates and speech timing in two closely related but rhythmically dissimilar varieties of British English, spoken by monolinguals from Leeds and Panjabi-English bilinguals from Bradford. The perception test with native expert listeners from the discipline showed that there was a systematic, albeit small, difference between spoken extracts produced by the two groups of speakers, which could be projected onto the continuum between syllable-timing and stress-timing. Speech from Bradford speakers was judged to be closer to the syllable-timed end of the continuum while speech from Leeds speakers received more stress-timed judgments, confirming previous work on these varieties (Heselwood and McChrystal, 2000) and further supporting the line of evidence that contact varieties of English often develop more syllable-timed than stress-timed characteristics (Deterding, 2001; Torgersen and Szakay, 2011; Low *et al.*, 2000).

Following the spirit of the *rhythm class* and *control and compensation* hypotheses, we then explored the role played by lengthenings (due to linguistic functions) and shortenings (due to coarticulatory forces) of consonants, vowels, syllables and rhythmic feet in the rhythmic typology. As far as the predictions put forward in 1.D are concerned, our findings offer partial support for both the

RCH- and the CCH-framed explanations of the bases for stress-timed/compensating rhythmic properties in Leeds and syllable-timed/controlling properties in the Bradford variety.

As expected according to the RCH-based predictions, positioning within the prosodic hierarchy had a stronger impact on the temporal properties of consonantal intervals in Leeds than in Bradford, where phrase-finally we found more cumulative lengthening as more consonants were added to the consonantal interval. These effects of the prosodic hierarchy were similarly well reflected in the results for syllable durations, where Leeds again showed more cumulative lengthening than Bradford phrase-finally, and syllables containing tense/diphthongal vowels also showed a greater sensitivity to phrasal position in Leeds than Bradford. In line with the CCH-based predictions, vowels in Leeds displayed more extensive compression effects as compared to Bradford: while Bradford vowels exhibited compression only in accented syllables, the effect was present regardless of the presence of prominence in Leeds. Echoing these results, accented and unaccented syllables presented similar degrees of lengthening due to an increasing number of consonants in Leeds, whereas in Bradford, unaccented syllables lengthened much more than accented ones.

Nevertheless, some of the predictions based on the two typologies were not supported by our data. We did not find a smaller number of syllables with reduced vowels in Bradford compared to Leeds. There were no cross-dialectal differences in the implementation of compression in complex consonant clusters, or in prominent syllables beginning rhythmic feet. Moreover, there was no increased amount of simpler consonant clusters in Bradford to support the articulatory perspective presented by the CCH. On the contrary, Bradford speakers unexpectedly produced a significantly higher number of syllables with a more complex structure. Neither did we find more instances of peripheral vowel qualities in weak syllables produced by Bradford speakers, as predicted by the RCH, and also shown in previous research on a syllable-timed contact variety of English (Low *et al.*, 2000). However, given that the perceptual differences between the two varieties investigated here are subtle and do not represent the prototypical opposite ends of the continuum between syllable-timing and stress-timing, the lack of supporting

evidence for each potentially relevant factor is perhaps not entirely surprising. Interestingly, dialect-specific effects involving compression and supporting the CCH were visible at the level of vowel timing, while prosodic hierarchy effects in support of the RCH primarily affected consonant durations. Vocalic and consonantal effects were mirrored and somewhat magnified at the level of syllables but disappeared completely at the level of rhythmic feet. The latter finding contrasts with the results of Krivokapić (2013) for Indian and American English, but aligns well with previous research involving a syllable-timed (Estonian) and a stress-timed (English) language which showed little difference with respect to foot-level timing (Asu and Nolan, 2006).

Taken together, these results reveal a more complex picture than the one suggested by either the RCH or the CCH. It seems that temporal flexibility resulting from both extensive lengthening *and* shortening may be one of the core features of a variety (and potentially, a language) classified as more stress-timed, when the influences from phonology and phonotactics are neutralized as in the present study. This finding helps to explain one of the reasons why cross-linguistic differences have repeatedly been evidenced by temporal variability indices, as represented by the traditional rhythm metrics. The output of the metrics applied to our dialectal data provided additional evidence for the same overall pattern, with Bradford speakers having lower variability scores and higher vowel-to-consonant ratios than Leeds speakers, once again in line with the syllable-timed/stress-timed distinction (cf. Deterding, 2001; Grabe and Low, 2002; Low *et al.,* 2000; Nolan and Asu, 2009; Ramus *et al*., 1999; White and Mattys, 2007). However, the timing patterns discussed above were not as straightforwardly reflected in the rhythm metrics as the general tendency confirmed here may imply.

First of all, the differences in the variability coefficients cannot have been triggered by the traditionally enumerated sources, vowel reduction and consonantal phonotactics (Dauer, 1983; Low *et al.,* 2000; Ramus *et al*., 1999), since the materials were identical across the dialects and we did not find any cross-dialectal differences in the percentages of syllables realized with complex structures or reduced vocalic nuclei, or of weak syllables with peripheral vowel qualities.

In the case of consonants, differences in the temporal demarcation of phrasal edges in Leeds vs. Bradford were fairly well picked up by the metrics. Given the higher number of phrase-medial than -final segments in the materials, it seems intuitive that the global measure Varco performed better than the local, iterative nPVI-metric whose output might be slightly diluted by phrase-medial similarities between the two varieties. However, there was a discrepancy between our results and previous findings which have argued for the application of raw, rather than rate-normalised metrics such as (ΔC and rPVI-C) to consonantal intervals, mainly because of their more successful performance in discriminating between languages of different rhythm classes (cf. Ramus *et al*., 1999; Grabe and Low, 2002, White and Mattys, 2007). None of the raw metrics showed a significant cross-dialectal difference here, while the rate-normalised metrics did, i.e. Varco-C and the (marginally significant) nPVI-C. It is interesting to note that rate normalization appears to remove phonologically conditioned variability such as found in cross-linguistic comparisons, yet seems necessary in order to observe phonetically conditioned variability such as occurs across dialects. These results add to the critical discussion of the rhythm metrics' reliability and, consequently, validity, extensively addressed by Knight (2011, cf. also Arvaniti, 2009).

The relationship between vocalic metrics and vowel duration patterns was quite weak. It was surprising that we found only a weak effect involving %V, not other metrics. The finding that Leeds differed from Bradford in having a higher degree of vocalic compression in unaccented syllables would rather lead one to expect a cross-dialectal difference in nPVI-V (because of more pronounced, local durational contrasts interspersing prosodic phrases in the Leeds data), which however was not significant here. As %V is not independent of consonantal properties, cross-dialectal results for consonant timing (such as a significantly higher proportion of consonantal syllable nuclei and a stronger reflex of prosodic edges in consonantal durations) may all have contributed to a significantly lower %V in Leeds. Evidently, these complex temporal relationships between consonantal and vocalic portions of the acoustic signal do not have an adequate expression in %V.

The performance of rhythm metrics proved particularly unsatisfactory at the level of syllable. Due to the cumulative effect of vowel and consonant durations, we found the strongest cross-dialectal timing differences at the syllable level, involving influences from vowel phonology, number of segments, prominence and phrasal position. These timing effects should have led to significant differences in Varco-S and nPVI-S scores, in particular, yet these failed to yield a significant output. The only level where we found a clear correspondence between the metrics' output and durational patterns was that of rhythmic feet, where we did not find any dialectal differences; but this result does not seem to provide very strong support for the validity of foot-level metrics.

The success of the CCI metric in expressing variable degrees of articulatory control vs. compensation behaviour across the two varieties should again be interpreted as rather mixed**.** While the absence of a significant effect for consonants and rhythmic feet was in line with the timing results, significantly different degrees of vocalic and syllabic compression in Leeds and Bradford drastically diverged from their rather uniform CCI-scores. At least in theory, the lack of syllabic compression (and of a significant CCI-score) could potentially stem from a vocalic compression and a consonantal lengthening cancelling each other out at the level of syllable. This, however, seems a rather unlikely explanation of the present result patterns, given that the timing effects on vowels and consonants are differentiated by prosody, i.e. compression of vowels interacts with prominence, lengthening of consonants interacts with phrasal boundaries (both effects are very well reflected in the results for syllable timing). Although the postulates of the CCH have enriched our understanding of the principles governing major cross-linguistic patterns of rhythm and timing, the CCI metric itself has not provided us with a useful tool for accessing these essential differences in coarticulation. Apparently, the weakness of the CCI score may stem from the fact that acoustic measurements of vocalic and consonantal intervals have been used to infer articulatory behaviour.

To summarise our main findings, even though numerically, we found the usual and expected patterns of metric outputs for a stress-timed as opposed to a syllable-timed variety (lower percentage of

vocalic intervals, higher variability of consonantal, vocalic, syllabic and foot intervals in Leeds), the differences were mostly slight or highly variable within speakers of each variety and therefore not significant. Most strikingly, substantial cross-dialectal timing effects failed to show up in syllabic metrics. Moreover, we failed to replicate the previously-found advantage for raw as opposed to rate-normalized consonant metrics. Despite the fact that many of the predictions of both typological views found support in our timing results, the rhythm metrics themselves failed to give an adequate picture of these data, despite the large size of our corpus in terms of materials if not number of speakers. This was true for the widely used Varco, PVI and %V scores as well as the more recently proposed and less well established CCI-metric. These results contribute to the growing body of evidence for a rather weak, or indeed absent, relationship between rhythm metrics and the timing phenomena that they are supposed to capture (e.g. Arvaniti, 2009; Barry *et al.*, 2009; Easterday *et al.*, 2011; Knight, 2011). Systemic differences in the phonologies of diverse languages are not well reflected in the metrics (Easterday *et al.*, 2011), and as we have seen in this study, nor are subtle (yet perceptible) realisational differences due to the dialect-specific implementation of the prosodic hierarchy and articulatory strategy.

B Sources of timing variation and origins of rhythmic typologies

The timing patterns identified in this study suggest that functional lengthening and coarticulatory shortening, proposed in the literature as two alternative underlying sources of linguistic rhythm typologies (Dauer, 1983; Low *et al.*, 2000; Ramus *et al*., 1999, for RCH and Bertini and Bertinetto, 2010, for CCH), are not mutually exclusive but seem to describe co-existing mechanisms of timing control, at least as far as rhythmically different varieties of English are concerned. A dialectal variety from the traditional stress-timed class (Leeds English) was found not only to exhibit stronger temporal reflexes of the prosodic hierarchy but also to be, in some respects, more compensating. That is, at least acoustically, it is characterized by an extensive implementation of both "temporal signatures of prosody" (lengthenings and shortenings, see Fletcher, 2010). In contrast, a variety that is perceptually more

syllable-timed (Bradford Panjabi bilingual English) employs shortenings and lengthenings to a somewhat lesser degree. While "controlling" behaviour has previously been attributed to a limited degree of segmental coarticulation (Bertini and Bertinetto, 2010), our findings suggest that the Bradford variety can be legitimately described as "controlling" also because of a higher amount of control over prosodic timing variability. As the postulates of the control and compensation hypothesis are based on articulation, it would be useful to confirm our acoustically-based conclusions by means of direct articulatory investigation. There is a contradiction between the theory and the practice of the CCH: on the one hand, the hypothesis is rooted in the idea that articulatory mechanisms can differ across languages, but on the other hand, only acoustic data are used to infer the underlying articulatory strategy.

Our results suggest the possibility that the rhythmic templates dominating typological views on linguistic rhythm since the last century (Abercrombie, 1967; Lloyd James, 1940; Pike, 1945) may have been based, to some extent, upon perceptual impressions arising from timing processes of more or less extensive lengthening and shortening, as well as upon variable degrees of vowel reduction and phonotactic complexity as previously suggested (Dauer, 1983; Low *et al.*, 2000). To what extent timing processes and phonological structure go hand in hand is to date rather poorly understood. A recent proposal suggests that 'stress-timing' and 'syllable-timing' may function as orthogonal variables of linguistic rhythm and co-exist at different levels of prosodic structure in the same language (Nolan and Asu, 2009). Against the background of our findings, it does not seem surprising that observations of durational variability have been at the centre of language-specific rhythm research (Deterding, 2001; Grabe and Low, 2002; Low *et al.*, 2000; Nolan and Asu, 2009; Ramus *et al.*, 1999; Rathcke and Smith, 2011; White and Mattys, 2007), although our results clearly indicate that it is incorrect to attribute this variability to the sole influence of language phonologies.

The existence (and potentially, the usefulness) of typological rhythm templates has sometimes been questioned in the recent debates about the nature of linguistic rhythm (among many, Arvaniti, 2009; Kohler, 2009). RCH-inspired research based on rhythm metrics certainly seems to have exhausted its

potential to teach us about the fundamentals of linguistic rhythm. The rather poor, or at least mixed, success of rhythm metrics as documented in our study as well as by previous research (e.g. Arvaniti, 2009; Barry *et al.*, 2009; Easterday *et al.*, 2011; Knight, 2011) can be partly explained by a limited understanding of how perceptual construction processes and articulatory motor constraints contribute to the phenomenon of linguistic rhythm, and how those mechanisms of spoken language may give rise to a rhythmic typology. The temporal structure of segmental, "micro-rhythmic" units and prosodic, "macro-rhythmic" units both play an important role in rhythm production and perception (cf. the cross-linguistic overview in Fletcher, 2010). If we assume that typological templates of some kind do exist, measuring durations of segmental intervals alone will never suffice to capture systematic tendencies in the acoustic substance of those templates. In an earlier study (Rathcke and Smith, 2011), we made an attempt to create a more linguistically grounded rhythm metric, the Multi-factorial Dispersion Coefficient. The preliminary version of the MDC combines structural lengthening (the product of lengthening coefficients) and variability (the root mean square of variability coefficients) on chosen linguistic factors. In particular, factors related to the demarcation of prominence and grouping, both often mentioned as core properties of perceived rhythm (e.g. Fletcher, 2010; Patel, 2008), can be straightforwardly integrated. The main purpose of the measure was to systematically capture cross-dialectal timing differences in those factors in the elegant and succinct way that is characteristic of rhythm metrics. The proposed MDC is still under development but constitutes the first step toward a metric which explores properties of speech timing beyond the RCH/CCH. It might offer a basis for testing the perceptual effects of various timing properties to elaborate a new concept of rhythmic typology. Future work would benefit from an interdisciplinary orientation, seeking to integrate insights from general psychological constraints (e.g. beat induction, subjective rhythmisation) and from linguistically sophisticated investigations of the "temporal signatures" of both phonological and prosodic structures in typologically diverse languages (cf. Arvaniti, 2009).

To conclude, we have presented evidence of variable implementations of the "temporal signatures of prosody" in different varieties of a language (cf. Fletcher, 2010: 579). Within-language comparison, keeping the phonological and semantic structure of materials constant, allowed us carefully controlled access to these variable prosodic implementations, which were found to differ subtly but significantly, in line with the subtle yet significant perceptual difference in rhythm between the varieties. Our study has further contributed to the much-debated question of linguistic rhythm and rhythmic typology (e.g. Arvaniti, 2009; Kohler, 2009) as well as to the critical evaluation of the explanatory adequacy of rhythm metrics (e.g. Barry *et al.*, 2009; Knight, 2011). We have demonstrated that research into rhythmic typologies benefits greatly by looking beyond the usually emphasized phonological factors and related low-level micro-timing variation, at the macro-timing effects due to the implementation of prosodic structure.

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council UK (RES-061-25-0302). The authors would further like to thank Francesco Li Santi for his contribution to the creation of the database, Ibrar Butt, Eve Carter, Leendert Plug and Barry Heselwood for supporting our data collection in Bradford and Leeds, Scott Jackson and Sam Miller for their advice on data analyses.

---

[1] In a subsequent development of the rhythm class hypothesis, a third group of languages was added to the typology, so-called 'mora-timed' languages like Japanese (see Ladefoged, 1975). Aside from being less well studied in comparison to syllable- and stress-timed languages, this group is not of primary relevance to the aim of this paper and will therefore not be discussed here.

[2] This view resonates with Grabe and Low's (2002) proposal that the distinction between stress- and syllable-timing is a gradient, not a categorical one, with languages aligning themselves along a continuum between the two rhythmic prototypes. Similarly, coupled oscillator models (e.g. Barbosa, 2007) assume competition between

a syllable and a foot oscillator which, depending on the language-specific strength of each oscillator, creates a gradient tendency towards syllable- or stress-timing.

REFERENCES

Abercrombie, D. (**1967**). *Elements of general phonetics.* (Edinburgh University Press, Edinburgh). 216 pages.

Arvaniti, A. (**2009**). "Rhythm, timing and the timing of rhythm," Phonetica **66**, 46-63.

Asu, E.L., and Nolan, F. (**2006**). "Estonian and English rhythm: a two-dimensional quantification based on syllables and feet," in *Proceedings of the 3rd Conference on Speech Prosody*, pp. 249-252.

Barbosa, P.A. (**2007**). "From syntax to acoustic duration: A dynamical model of speech rhythm production," Speech Commun. **49(9)**, 725–742.

Barry, W., B. Andreeva, B., and Koreman, J. (**2009**). "Do rhythm measures reflect perceived rhythm?" Phonetica **66**, 78-94.

Bertini, C., and Bertinetto, P.M. (**2010**). "Towards a unified predictive model of Natural Language Rhythm," in *Prosodic Universals. Comparative Studies in Rhythmic Modeling and Rhythm Typology*, edited by M. Russo (Roma: Aracne), pp. 43-78.

Browman, C. P., and Goldstein, L. (**1989**). "Articulatory gestures as phonological units," Phonology **6**, 201-251.

Byrd, D. (**1995**). "C-centers revisited," Phonetica **52**, 285-306.

Byrd, D. (**1996**). "Influences on articulatory timing in consonant sequences," J. Phonetics **24**, 209-244.

Classe, A. (**1939**). *The rhythm of English prose.* (Oxford: Blackwell). 138 pages.

Dauer, R. M. (**1983**). "Stress timing and syllable-timing reanalyzed," J. Phonetics **11**, 51-62.

Dellwo, V., and Wagner, P. (**2003**). "Relations between language rhythm and speech rate," in *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 471–474.

Deterding, D. (**2001**). "The measurement of rhythm: a comparison of Singapore and British English," J. Phonetics **29**, 217-230.

Easterday, S., Timm, J., and Maddieson, I. (**2011**). "The effects of phonological structure on the acoustic correlates of rhythm," in *Proceedings of the 17th International Congress of Phonetics Science*, pp. 623-626.

Fletcher, J. (**2010**). "The prosody of speech: timing and rhythm," in *The handbook of phonetic sciences*, edited by W.J. Hardcastle, J. Laver and F.E. Gibbon (Oxford: Blackwell), pp. 521-602.

Fowler, C. (**1977**). *Timing control in speech production* (Bloomington: Indiana University Linguistics Club). 186 pages.

Fowler, C. (**1981**). "A relationship between coarticulation and compensatory shortening," Phonetica **38**, 35-50.

Grabe, E. (**2001**). Documentation for the IViE corpus. Retrieved from http://www.phon.ox.ac.uk/files/apps/old_IViE/documentation.html (Last viewed 03/03/2015).

Grabe, E., Low, E.L. (**2002**). "Durational variability in speech and the Rhythm Class Hypothesis," in *Papers in Laboratory Phonology 7*, edited by N. Warner and C. Gussenhoven (Berlin: Mouton de Gruyter), pp. 515-546.

Harrington, J. (**2010**). *Phonetic analysis of speech corpora* (Oxford: Wiley-Blackwell). 424 pages.

Hayes, B. (**1995**). *Metrical stress theory: principles and case studies.* (Chicago: The University of Chicago Press). 472 pages.

Heselwood, B., and McChrystal, L. (**2000**). "Gender, accent features and voicing in Panjabi-English bilingual children," Leeds Working Papers in Linguistics and Phonetics **8**, 45-70.

Katz, J. (**2012**). "Compression effects in English," J. Phonetics 40, 390-402.

Knight, R-A. (**2011**). "Assessing the temporal reliability of rhythm metrics," J. Intern. Phonet. Assoc. **41(3)**, 271-281.

Kohler, K.J. (**2009**). "Rhythm in speech and language: A new research paradigm," Phonetica **66**, 29-45.

Krivokapic, J. (**2013**). "Rhythm and convergence between speakers of American and Indian English," Laboratory Phonology **4(1)**, 39-65.

Ladefoged, P. (**1975**). *A course in phonetics* (New York: Harcourt Brace Jovanovich). 296 pages.

Lehiste, I. (**1977**). "Isochrony reconsidered," J. Phonetics **5(3)**, 253-263.

Lloyd James, A. (**1940**). *Speech signals in telephony.* Cited in Abercrombie (1967; p.171).

Low E.L., Grabe E., and Nolan, F. (**2000**). "Quantitative characterisations of speech rhythm: syllable-timing in Singapore English," Language and Speech **43(4)**, 377-401.

Nolan, F., and Asu, E.L. (**2009**). "The pairwise variability index and co-existing rhythms in language," Phonetica **66**, 64-77.

Patel, A. (**2008**). *Music, language and the brain* (Oxford: Oxford University Press). 526 pages.

Peterson, G., and Lehiste, I. (**1960**). "Duration of syllabic nuclei in English," J. Acoust. Soc. of Am. **32**, 693–702.

Pike, K. L. (**1945**). *The intonation of American English.* (University of Michigan Press, Ann Arbor). 200 pages.

Prieto, P., Vanrell, M.d.M., Astruc, L., Payne, E., and Post, B. (**2012**). "Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish," Speech Commun. **54(6),** 681–702.

Ramus, M. Nespor, and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73(3),** 265–292.

Rathcke, T., and Smith, R. (**2011**). "Exploring timing in accents of British English," in *Proceedings of the 17th International Conference on Phonetic Sciences,* pp. 1666-1669.

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

Selkirk, E.O. (**1981**). "English compounding and the theory of word-structure," in *The Scope of Lexical Rules*, edited by M. Moortgat, H. Van der Hulst and T. Hoestra (Dordrecht: Foris), pp. 229-277.

Torgersen, E., and Szakay, A. (**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),** 165-174.

Vayra, M., Fowler, C., and Avesani, C. (**1987**). "Word-level coarticulation and shortening in Italian and English speech," Studi di Grammatica Italiana **13**, 249-269.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (**2010**). "How stable are acoustic metrics of contrastive speech rhythm?" J. the Acoustical Society of America **127**, 1559–1569.

White, L., and Mattys, S. L. (**2007**). "Calibrating rhythm: First language and second language studies," J. Phonetics **35**, 501–522.

White, L., and Turk, A. (**2010**). "English words on the Procrustean bed: Polysyllabic shortening reconsidered," J. Phonetics 38, 459–471.

White, L., Payne, E., and Mattys, S.L. (**2009**). "Rhythmic and prosodic contrast in Venetian and Sicilian Italian," in *Phonetics and Phonology: Interactions and Interrelations*, edited by M. Vigario, S. Frota and M.J. Freitas (Amsterdam: John Benjamins), pp. 137-158.

Tables

Table 1: Mean values (and standard deviations) of four vocalic metrics for Leeds and Bradford varieties.

	%V	Varco-V	nPVI-V	CCI-V
Leeds	38.7 (0.7)	0.60 (0.04)	64.1 (4.2)	47.3 (7.2)
Bradford	40.2 (1.9)	0.57 (0.05)	61.1 (2.4)	48.6 (5.0)

Table 2: The best-fit model for the duration of vocalic intervals.

Predictors	Df	χ^2	p
*phrasal position*accentuation*	2	26.7	<0.001
*vowel quality*phrasal position*number of segments*	4	45.9	<0.001
*dialect*accentuation*number of segments*	1	6.5	0.01

Table 3: Mean values (and standard deviations) of five consonantal metrics for Leeds and Bradford varieties.

Dialect	ΔC	Varco-C	rPVI-C	nPVI-C	CCI-C
Leeds	57.3 (5.5)	0.53 (0.01)	66.4 (7.3)	62.1 (2.6)	38.4 (3.7)
Bradford	54.4 (5.0)	0.50 (0.02)	63.0 (6.4)	59.3 (2.3)	35.9 (2.3)

Table 4: The output of the model fitting procedure for the duration of consonantal intervals.

Predictors	Df	χ^2	p
*number of consonants*phrasal position*structural position*	2	42.0	<0.001
*number of consonants*structural position*accentuation*	1	5.6	< 0.05
*dialect*number of consonants*phrasal position*	2	15.7	<0.001
*number of consonants*phrasal position*accentuation*	2	12.8	<0.01
*dialect*phrasal position*structural position*	2	19.7	<0.001

Table 5: Mean values (and standard deviations) of three metrics applied to syllables (S) and feet (F) in Leeds vs. Bradford.

	Varco-S	nPVI-S	CCI-S	Varco-F	nPVI-F	CCI-F
Leeds	0.48 (0.02)	55.1 (2.2)	33.3 (4.6)	0.40 (0.02)	46.3 (2.9)	78.7 (7.6)
Bradford	0.46 (0.03)	55.0 (1.0)	33.0 (2.7)	0.38 (0.01)	44.4 (2.5)	82.1 (5.6)

Table 6: The best-fit model for the duration of syllable intervals.

Predictors	Df	χ^2	p
*dialect*phrasal position*number of segments*	2	21.6	<0.001
*dialect*vowel quality*number of segments*	2	14.2	<0.001
*number of segments*vowel quality*phrasal position*	4	17.2	<0.01
*dialect*accentuation*number of segments*	1	19.5	<0.001
*dialect*phrasal position*vowel quality*	4	31.7	<0.001
*dialect*accentuation*phrasal position*	2	20.2	<0.001

Table 7: The best-fit model for the duration of foot intervals.

Predictors	Df	χ^2	p
vowel quality	1	9.5	<0.01
*accentuation*phrasal position*number of syllables*	6	351.1	<0.001