RHYTHM CLASS PERCEPTION BY EXPERT PHONETICIANS

Tamara V. Rathcke[1] and Rachel H. Smith[2]

[1]University of Kent, [2]University of Glasgow
[1]T.V.Rathcke@kent.ac.uk, [2]Rachel.Smith@glasgow.ac.uk

ABSTRACT

This paper contributes to the recent debate in linguistic-phonetic rhythm research dominated by the idea of a perceptual dichotomy involving "syllable-timed" and "stress-timed" rhythm classes. Some previous studies have shown that it is difficult both to find reliable acoustic correlates of these classes and also to obtain reliable perceptual data for their support.

In an experiment, we asked 12 British English phoneticians to classify the rhythm class of 36 samples spoken by 24 talkers in six dialects of British English. Expert listeners' perception was shown to be guided by two factors: (1) the assumed rhythm class affiliation of a particular dialect and (2) one acoustic cue related to the prosodic hierarchy, namely the degree of accentual lengthening.

We argue that the rhythm class hypothesis has reached its limits in informing empirical enquiry into linguistic rhythm, and new research avenues are needed to understand this multi-layered phenomenon.

Keywords: rhythm class, rhythm perception, expert phonetician, dialects of British English

1. INTRODUCTION

The long-standing typology of linguistic rhythm assumes that the perception of rhythm in languages can be mapped onto two templates, or rhythm classes, named "syllable-timing" and "stress-timing" [1], with e.g. Romance languages belonging to the former and Germanic languages to the latter category. Early proposals believed that rhythm classes were underpinned by isochronous timing of either syllables or stresses, giving rise to the two perceptual templates [1]. However, any attempt to find evidence for isochronous intervals in speech has failed [2,5].

In consequence, the idea of isochrony as the acoustic basis for the rhythm class dichotomy was abandoned and soon replaced by the assumption that rhythm class perception stems from language-specific phonology, specifically the presence of reduced vowels and the complexity of consonant phonotactics [5,15]. Accordingly, "stress-timing" is related to a high level of vowel reduction in unstressed syllables and the presence of complex consonant clusters while "syllable-timing" goes hand-in-hand with a simple syllable structure and the absence of vowel reduction.

Subsequently, so-called rhythm metrics were developed to capture the temporal manifestations of these linguistic properties [15]. The most popular and frequently used metrics include: (1) a vowel-to-consonant ratio, expressed as a percentage (%V, [15]); (2) the variability coefficient Varco, calculated as the standard deviation divided by the mean duration of a consonantal or a vocalic interval [6]; (3) the pairwise variability index PVI, calculated as the mean temporal distance between pairs of successive consonantal or vocalic intervals (raw or normalized to the mean duration of each pair of intervals [12]).

It has been assumed, and frequently shown, that "syllable-timing" produces high vocalic proportions of utterances and low variability scores while "stress-timing" usually manifests itself in lower %V and higher consonantal and vocalic variability (among many, [6,12,14,15,19]). Even dialects of a language have been classified within this conceptual framework [12,20]. In particular, "syllable-timing" has often been associated with contact varieties of English: Multicultural London English and also varieties spoken in Singapore and by Punjabi-English bilinguals in Yorkshire have been previously classified as "syllable-timed" ([11,12,19]).

Despite the initial success of metrics, their ability to provide the much needed support to the idea of the two distinct rhythm classes has recently been questioned after a series of rather critical findings showing that, e.g. a given language may be classified differently depending on the type of metrics used [8]; differences in materials, speaking styles or speech rates introduce larger changes in metric scores than differences across languages do [2]; and vocalic metrics do not always straightforwardly pick up on vowel reduction present in a language [7].

Following these criticisms, new proposals suggest that rhythm class may be grounded in prosodic timing alternations, i.e. variable degrees of lengthening at different levels of the prosodic hierarchy specific to a language or a dialect [17,20]. Languages or dialects with "stress-timing" are more likely to employ a strong temporal demarcation of accentuation and phrasing than languages or dialects with "syllable-timing".

However, the crucial piece of evidence in the rhythm debate has to be sought in perception. If it is difficult or even impossible to pin down the acoustic substance of a rhythm concept, it might be not the phonetic measures that cause the problem but the concept itself.

So far, the strongest evidence in support of the dichotomy comes from a series of language discrimination experiments with linguistically impoverished samples [16]. The finding that languages from the same rhythm class are poorly discriminated in contrast to languages from the different classes, however, seems to be an artefact of speech rate differences in the stimuli [3].

In identification experiments, on the other hand, naïve participants cannot reliably use the two categories, either in a metalinguistic or in a formal experimental setting [3,13], and even phonetically trained listeners find the task difficult and do not produce consistent patterns [13]. Miller [13] hypothesised that the diverse linguistic background of the expert listeners may have led them to attend to different acoustic cues and could thus explain the lack of agreement in perceptual judgments.

The present study investigates if there is a consensus among specialists with a homogeneous linguistic background, judging naturalistic stimuli produced in different dialects of their native language, and aims at answering the following questions:

(1) Can expert phoneticians reliably use "syllable-timing" and "stress-timing" as perceptual anchor points when categorising regional and ethnic varieties of their native language (classified as belonging to different rhythm categories in previous studies)? If, as we assume, this is indeed the case, the follow-up question arises:

(2) What is the basis for experts' perceptual judgements? There are potentially two sources of listeners' informed judgement. First, there may be some acoustic features they rely on. Given the recent critical work [2,7], we could expect traditional rhythm metrics to serve as rather poor predictors of perceptual categorisation. Instead, speech rate [3], demarcation of the prosodic hierarchy [17,20] and spectral changes in vowels due to the presence or absence of phrasal prominence [5,15] were deemed of particular importance in this study.

Alternatively, experts' perception might be guided by their *apriori* knowledge of a putative rhythm class affiliation leading to top-down categorization behaviour in a formal perception test setting.

2. METHOD

2.1. Listeners

Thirty native English phoneticians from the UK with expert knowledge of the dichotomy were approached by email. Fifteen of them responded. The dataset of this paper is based on judgments by 12 participants (two datasets could not be analysed due to technical issues; one phonetician admitted to having difficulties with the perception of prominence and doubts about the rhythm class concept in particular).

2.2. Speakers

The following 6 regional and ethnic accents were studied here: Belfast, Bradford (Panjabi-English bilingual speakers), Cambridge, Leeds, Newcastle and London (speakers of West Indian descent). Two of these accents were expected to elicit more syllable-timed responses, Bradford and London.

2.3. Stimuli

The stimuli for this investigation were taken from two corpora of read speech, containing recordings of British English speakers with different regional accents [17, 18]. From these two corpora, fairy-tale passages ("The Princess and The Pea", "Cinderella") and "The Sailor" passage were selected (the latter was composed to provide the widest possible range of potential differences between broad regional accent types across the UK [4]).

Thirty-six short samples (6 seconds on average, ranging from minimally 4 to maximally 8 seconds) were used. The samples consisted of 13 coherent passage extracts, read fluently in two different accents. Eleven extracts were represented both in a putatively "syllable-timed" (Bradford, London) and a "stress-timed" accent (Belfast, Leeds, Newcastle). Two extracts were represented in two "stress-timed" accents only (spoken by a male and a female talker from Leeds or Cambridge respectively).

Overall, data from 24 speakers (12 f) were used in the perception experiment. All samples were low-pass filtered at 1 kHz to make the sound quality comparable across the two corpora.

2.4. Procedure

The participants were sent a PowerPoint presentation containing the stimuli and brief instructions. Each stimulus was paired with a continuum (depicted by a 120 mm long straight line) spanning the two poles, labelled "strongly syllable-timed" on the left and "strongly stress-timed" on the right (cf. [8]). 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.

There were 3 randomisation lists. Although we approached an equal number of listeners per list, the final dataset contained an unequal number of responses for each order of stimulus presentation (6 listeners for list 1, 5 listeners for list 2 and 1 listener for list 3).

2.5. Data preparation

Participant responses were measured in mm and transformed to a z-score. The test stimuli themselves were prepared for acoustic analyses in Emu/R. Consonant and vowel intervals were segmented. Vowels were further specified as accented or unaccented, phrase-medial or -final.

Subsequently, durations of consonantal and vocalic intervals were measured and %V, rPVI-C, nPVI-V, Varco-V and Varco-C metrics calculated, along with the degree of phrasal and accentual lengthening as well as speech rate (syll/sec). To fully represent the phrasal and accentual structure in the stimuli, we included the number of pauses and prosodic breaks within each sample, the average phrase duration in syllables, and the ratio of the number of accented to unaccented syllables.

Pitch and spectral measurements were taken at vowel midpoints. Vocalic spectra were represented by the first four DCT coefficients [9]. Pitch measurements included overall pitch range and mean f0-difference between accented and unaccented vowels in semitones whereas the information about vowel spectra comprised of Euclidean distances between the means of four DCT coefficients of accented and unaccented vowels.

2.6. Statistical analyses

All analyses were conducted in R using emu, lme4, lmerTest and diptest libraries. First, Hartigans' dip test statistic [10] was run on the perception data to confirm that listeners had a clear concept of two different rhythm classes, associated with either end of the continuum. The output of this statistic (dn) indicates how much a distribution drops below an expected unimodality curve (i.e. a larger dn-score corresponds to a less unimodal distribution).

Factor analysis was run on the fourteen parameters of interest (metrics, prosodic and spectral measures). Four of the resulting factors were analysed with respect to their loadings. Lastly, the acoustic parameters most representative of each factor (i.e. with the highest loading in the corresponding factor, and the lowest loading elsewhere) were selected and, in addition to rhythm class, used as predictors in a mixed-effects regression. Z-scored perceptual responses served as the dependent variable. Listener, speaker, extract and the order of presentation were specified as random effects.

3. RESULTS

In the very first run of statistical analyses, perceptual responses to all experimental stimuli were analysed according to the protocol described above. No significant patterns were found in the responses. In the second run, we excluded judgments of four experimental samples that were presented only in two "stress-timed" accents, Leeds and Cambridge, and lacked a syllable-timed pair accent. The remaining dataset contained pairs of stimuli that were presented in both a syllable- and a stress-timed accent, and consisted of 32 samples in total. Significant patterns started to appear in this slightly, but meaningfully, reduced sample.

Figure 1 presents a nonparametric density plot of the z-scored perception data. Hartigans' dip test statistic showed that the distribution of the responses differed significantly from a unimodal distribution at $p < 0.01$ ($dn = 0.021$), providing evidence for a two-way categorical judgment profile in these data.

Figure 1: Density curves fit for the z-transformed response distributions (N=384).

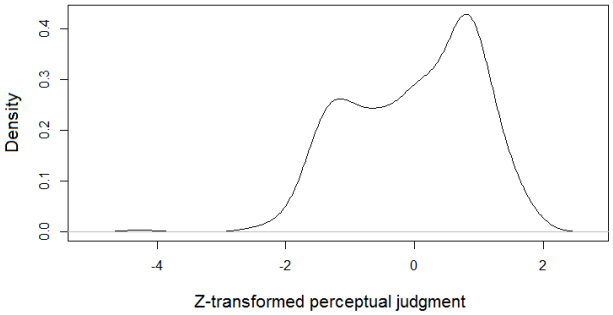

Table 1 summarizes the output of the factor analysis. All parameters with a loading higher than 0.5 are listed for each factor. Accordingly, Factor 1 identified features related to phrasal structure, Factors 2 and 4 picked up on features of phrasal prosody (note that Varco-V had high loadings on both factors), Factor 3 was dominated by the pairwise metrics and also loaded by speech rate. The spectral and f0-distance between accented and unaccented vowels, %V and Varco-C did not load highly on any of the factors.

A linear mixed-effects model was carried out with the putative rhythm class of the sample, number of phrases, nPVI-V, accentual and final lengthening as five predictors, and with z-scored listener responses as the dependent variable. Note that most of the predictors (including the fluency of the reading, timing phenomena, ethnic origin of the speaker) also cropped up in the informal comments of the experts

after they finished the experiment. The analysis revealed two significant effects.

Table 1: Four factors with their parameter loadings and percentage of variance reduced.

N	variance	parameters	load
#1	21.1%	*N of phrases*	0.961
		N of pauses	0.902
		N of syllables/phrase	-0.847
#2	13.3%	*Accentual lengthening*	0.773
		F0 range	0.585
		Varco-V	0.561
#3	12.5%	*nPVI-V*	0.739
		rPVI-C	0.742
		Speech rate	-0.512
#4	11.2%	*Final lengthening*	0.819
		Varco-V	0.766

The mixed-effects statistic revealed that in this dataset, the degree of accentual lengthening (χ^2=6.69, p<0.01) and the putative rhythm class affiliation (χ^2=9.12, p<0.01) helped to obtain an optimal model fit.

Stimuli with little accentual lengthening received scores closer to the "syllable-timed" end of the continuum while an increase in lengthening moved the scores closer to the opposite end (t(80.5)=2.58, p<0.05). Samples taken from London and Bradford speakers' readings received judgements that were approximately 0.6 z-scores closer to the "syllable-timed" end of the continuum than samples from Leeds, Belfast or Newcastle speakers (t(16.7)=3.27, p<0.01).

In terms of the actual placement along the continuum, these results corresponded to 56% of the scale for Bradford, London and 68% for Leeds, Belfast and Newcastle (where 0% means "strongly syllable-timed" and 100% corresponds to "strongly stress-timed" judgments). Overall, raw data showed a bias to locate all samples of this study towards the stress-timed end of the continuum (with the grand mean of 62%). However, this bias disappeared completely in the z-scored data (with means of -0.28 and 0.28 for the two groups of accents).

4. DISCUSSION

This study set out to investigate whether phoneticians with a specialist knowledge of the rhythm class dichotomy and from a homogeneous linguistic background would be able to reliably classify dialects of their language as either "syllable-timed" or "stress-timed". On the basis of the results we can conclude that this was indeed possible. The listeners showed a clearly bimodal distribution of their responses to the stimuli, with a slight bias toward the stress-timed end of the continuum in non-normalised data.

But which acoustic properties of the samples guided the experts' perception? We tested the predictive power of phrasing (exemplified through the number of phrases per sample), prosodic shape of the phrases (the degree of accentual and phrase-final lengthening) and variability of successive vocalic intervals (correlated with speech rate in these data, see Table 1). Only prominence-related cues were relevant to the task; no other features played a role in this test. To some extent, this result supports traditional accounts of rhythm which highlight the importance of prominence for rhythm perception, e.g. [2]. However, it was not the timing of prominence occurrence that was found to have an impact on rhythmic categorisation in this task, but the degree of prominence demarcation, resonating with the concept of a "prominence gradient" proposed in [14]: expert listeners (with a British English background) expected a shallow gradient in "syllable-timing" and a steeper one in "stress-timing".

Crucially, the results also suggest that experts' *apriori* knowledge about the rhythm class affiliation of regional and, most of all, ethnic dialects of their native language had a major influence on their perceptual judgments. This result is further supported by the fact that, unless the dataset was balanced with respect to pairings of "syllable-timed" vs. "stress-timed" samples, none of the factors showed sufficient power to reliably explain the variance in the dataset. If the subject-specific mean was skewed toward the "stress-timing" end of the continuum by a higher number of "stress-timed" samples, the distribution of responses appeared less clearly structured. Given that the number of listeners participating in the experiment was relatively low and the perceptual effects were rather small across the board, statistical analyses might have been underpowered, thus preventing us from seeing clear effects of phonetic parameters in the overall dataset. Increasing the number of participants is unlikely to completely remove the top-down effect. Rather, a replication of the study with non-native experts could potentially show increased use of acoustically guided, bottom-up perceptual strategies, coming, however, at a cost of language-specific preferences for the encoding of rhythm [13].

The rhythm class hypothesis has now seen three waves of searching for an acoustic substantiation of the dichotomy – isochronous timing, segmental phonology and most recently prosodic timing. All of these parameters may have a tight relationship to the multi-layered phenomenon of rhythm across the vast diversity of languages and speaking styles. But the rhythm class hypothesis itself has reached its limits in informing linguistic-phonetic enquiry [2,5,14].

5. REFERENCES

[1] Abercrombie, D., 1967. *Elements of general phonetics.* Edinburgh: Edinburgh University Press.

[2] Arvaniti, A., 2009. Rhythm, timing and the timing of rhythm. *Phonetica 66,* 46-63.

[3] Arvaniti, A., Rodriquez, T. 2013. The role of rhythm class, speaking rate and F0 in language discrimination. *Laboratory Phonology 4,* 7-38.

[4] Barry, W.J., Hoequist, C.E., Nolan, F.J. 1989. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language 3,* 355-366.

[5] Dauer, R. M., 1983. Stress timing and syllable-timing reanalyzed. *Journal of Phonetics 11,* 51-62.

[6] Dellwo, V., Wagner, P., 2003. Relationships between rhythm and speech rate. *Proc. 15th ICPhS* Barcelona, 471-474.

[7] Easterday, S., Timm, J., Maddieson, I., 2011. The effects of phonological structure on the acoustic correlates of rhythm. *Proc. 17th ICPhS* Hong Kong, 623-626.

[8] Grabe, E., Low, E.L. 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. *Papers in Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 515-546.

[9] Harrington, J. 2010. *Phonetic analysis of speech corpora*. Oxford: Wiley-Blackwell.

[10] Hartigan, J. A., Hartigan, P. M. 1985. The dip test of unimodality. *Annals of Statistics 13,* 70–84.

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

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

[13] Miller, M. 1984. On the perception of rhythm. *Journal of Phonetics 12,* 75-83.

[14] Nolan, F., Asu , E.L. 2009. The pairwise variability index and coexisting rhythms in language. *Phonetica 66,* 64-77.

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

[16] Ramus, F., Dupoux, E., Mehler, J. 2003. The psychological reality of rhythm class: Perceptual studies. *Proc. 15th ICPhS* Barcelona, 337-340.

[17] Anonymous. Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies. Submitted.

[18] The IViE corpus. English intonation in the British Isles. http://www.phon.ox.ac.uk/files/apps/IViE/

[19] Torgersen, E., 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.

[20] White, L., Payne, E., Mattys, S.L. 2009. Rhythmic and prosodic contrast in Venetan and Sicilian Italian. M. Vigario, S. Frota and M.J. Freitas (eds.), *Phonetics and Phonology: Interactions & Interrelations.* Amsterdam: John Benjamins, 137-158.