San Diego Linguistics Papers **4** *(2013) 28-52*

Clusters and Classes in the Rhythm Metrics

Russell Horton and Amalia Arvaniti

Abstract. The classic formulation of the rhythm class hypothesis holds that human languages are either stress-timed, with roughly equal duration of periods in between stresses, or syllable-timed, with all syllables tending toward equal duration. This strict isochrony failed to be confirmed empirically, and various metrics have been proposed to quantify the typology in terms of durational variation of segments instead. We show that a Naive Bayes classifier and a graph-based clusterer can use certain metrics to discriminate between stress-timed English and German and syllable-timed Italian and Spanish with greater than 80% accuracy. They cannot, however, consistently and strongly predict rhythm types for sentences from the unclassified languages Greek and Korean, indicating that the rhythm typology may not generalize beyond the prototypical exemplars of each class. Although the purported success of the metrics in modeling the rhythm typology has been used as a proof of the existence of the typology itself, this is unwarranted, as the main argument that the metrics are valid measures of rhythm is the fact that they partially accord with the predictions of the rhythm typology. Further, the success of the metrics is in proportion to the degree that they correlate with tempo, suggesting that they may merely reflect the reduced durational variability of segments at the high speaking rates that characterize syllable-timed languages, rather than reflecting inherent differences in rhythm class timing. Although the clusters generated on the basis of the metrics appear to replicate the typology fairly well, this is due to partially overlapping ranges between languages and not strict separation. We show that the actual distributions of the metric values are incompatible with the strong claims made for the categorical, universal and fundamental nature of rhythm in speech.

1 Introduction and Background

1.1 Origins of the Rhythm Typology

In his 1940 primer on telephony for a war-time audience, Arthur Lloyd James stresses the importance of rhythm for effective communication. Speech rhythms, he asserts, are of one of two kinds: the "morse-code" rhythm of languages like English, and the "machine-gun" rhythm of languages like French (Lloyd James, 1940). Five years later, Kenneth Pike formalizes this distinction, devoting three pages in "The Intonation of American English" to a topic he calls "Simple Rhythm Units (Stress-Time and Syllable-Time)"(Pike, 1945). In those pages, he sets out a theory of rhythm whose broad outlines remain influential today, even if many of its core precepts have had to be abandoned, "simplicity" first among them.

Pike describes English as being structured around rhythm units, each containing a single strongly stressed segment, and each of roughly equal duration. Due to this principle of isochrony, the main stresses determine

*This paper builds on work from my final paper for LIGN201, co-authored with Cody Brimhall and Emily Morgan, which was presented as a poster at the Acoustic Society of America in 2010, and from an associated paper draft co-authored with Brimhall, Morgan and Amalia Arvaniti.

the timing pattern of speech generally, and hence English is a *stress-timed* language. In contrast, a language such as Spanish is characterized by the equal duration of each syllable, regardless of stress, and can therefore be called *syllable-timed*.

Two decades later, Abercrombie greatly expanded Pike's ideas about rhythm, binding it in an intimate, physical manner to both production and perception. "[S]peech rhythm is essentially a muscular rhythm", he writes, and as hearers, "we *feel* it, entering empathetically into the movements of the speaker", with such "phonetic empathy" playing a major role in comprehension (Abercrombie, 1967). He also surpasses Pike, whose chief concern is with the characterization of English, in asserting the categoricality and universality of the rhythm typology. For Abercrombie, every language in the world is either stress-timed or syllable-timed, and crucially not both[1]. A less populous third class of mora-timed languages, typified by Japanese, has also been proposed.

Abercrombie's views represent the rhythm typology hypothesis at its strongest: rhythm is isochronic, universal, categorical, and critical for production and comprehension. The amount of subsequent work on the rhythm typology is hardly surprising if one takes Abercrombie's view that "[i]t is probable that the rhythm of a language is one of the most fundamental things about it, in the sense that it is one of the earliest things learnt by the infant, and perhaps the most difficult thing for the adult speaker to modify, when he wants to learn to pronounce a foreign language."

1.2 Support for the rhythm typology

Infant studies have lent support to Abercrombie's view of rhythm as a linguistic fundamental that is active early in life, with some results suggesting a innate sensitivity to rhythm and the rhythm classes. Nazzi, Bertoncini, and Mehler (1998) have shown that newborn French infants are able to discriminate between low-pass filtered speech from English and Japanese, which are taken to be stress-timed and mora-timed respectively, but not between English and Dutch, which are both stress-timed. Speech was filtered to remove all signals above 400-Hz in an attempt to remove segmental information and retain prosodic cues, to ensure that discrimination was on the basis of rhythm and not other speech properties.

The presence of such discriminative ability in newborns is taken as validation of the rhythm typology, and is also seen as suggesting that rhythm is an innate and fundamental property of human language with a strong role in language acquisition and processing. It has been speculated that the process of speech segmentation varies according to rhythm class, with stress-timed languages basing segmentation on stress patterns, and syllable-timed languages using the syllable as the fundamental segmentation unit. The innate appreciation of rhythm, allowing the infant to quickly determine the rhythmic type of her native tongue, is said to provide the "prosodic bootstrapping" necessary to segment speech into the proper units, and eventually to discern word boundaries (Nazzi & Ramus, 2003).

Modeling studies also provide some support for the rhythm class hypothesis. Recurrent neural network models have been used to discriminate English from Japanese (Dominey & Ramus, 2000). Their neural model's behavior is claimed to accurately simulate the rhythm sensitization behavior shown by the infants in Nazzi et al. (1998). Discrimination between languages of different rhythm classes has also been achieved using a complex Gaussian mixture model, with data derived from the durations of vocalic and consonantal intervals (Rouas, Farinas, Pellegrino, & André-Obrecht, 2005).

Abercrombie's supposition that rhythm is difficult to modify when learning a foreign language has also been empirically investigated. White and Mattys (2007) found some evidence that L2 speakers use a rhythm that is intermediate between the L2 language and their native tongue.

[1] As Abercrombie points out, the rhythm types are incompatible by definition, as the adaptation required to maintain even timing of syllables forces stresses to be produced unevenly, and vice versa.

1.3 Isochrony refuted

Some of the research Pike and Abercrombie inspired, however, proved to be incompatible with one of the central tenets of the rhythm typology, that of isochrony. The idea of isochrony had been largely based on qualitative judgements of how various languages struck the ear, such as Lloyd James' (1940) "morse-code" (stress-timed) and "machine-gun" (syllable-timed) characterizations. Further evidence was adduced from the poetic traditions in those languages, the idea being that the inherent rhythmic properties of a language are "[c]ontrolled strictly and mechanically in poetry", as Pike put it, and hence made explicit.

Empirical investigation, however, failed to bear out the claim that stresses occur more evenly in "stress-timed" languages such as English, or that syllables are of more even duration in "syllable-timed" languages like Spanish (e.g. Dauer, 1983; Bertrán, 1999; for reviews of such work see Bertinetto, 1989; Kohler, 2009). Due to the lack of empirical support, the appealing idea that a language's rhythm could be described by the equal duration of one of two simple rhythmic units was largely abandoned.

1.4 The Rhythm Metrics

Without the underpinning of isochrony, but still with a strong sense that the rhythmic differences expressed in the typology were real and important, researchers turned to quantitative methods to attempt to rehabilitate the theory. These proposed metrics turn a perceived liability – the fact that durations of rhythm units do not appear to be equal – into an asset, by measuring the relative durations and durational variability of these intervals. Stress-timed languages are characterized by a higher frequency of complex onsets and codas, leading to a greater variety of syllable structures. This suggests that they should exhibit greater variability in the durations of their consonantal and vocalic intervals than is found in syllable-timed languages, which have simpler, more consistent syllable structures. Stress-timed languages also exhibit more extensive vowel reduction than syllable timed languages. Reduced vowels, combined with more consonantal material in each syllable due to more consonant clusters, leads to the overall proportion of the speech signal that is occupied by vocalic intervals being lower in stress-timed languages than in syllable-timed languages. The aim of the rhythm metrics was to capture these timing differences, establishing the empirical support for the rhythm typology that isochrony had failed to provided.

Ramus, Nespor, and Mehler (1999) proposed three different metrics, based on the segmentation of the speech stream into alternating, complementary vocalic and consonantal intervals. $\%V$ is simply the percentage of the speech sample occupied by vocalic intervals, as calculated in (1):

$$\%V = \frac{100}{d_s} \sum_{k=1}^{m_v} d_k \tag{1}$$

where m_v is the number of vocalic intervals in the utterance, d_k is the duration of the kth vocalic interval, and d_s is the duration of the entire utterance. The complementary metric %C would simply be $100 - \%V$.

The additional metrics proposed by Ramus et al. are ΔC and ΔV, the standard deviations of the durations of the consonantal and vocalic intervals, respectively. ΔC is computed as follows:

$$\Delta C = 100 \sqrt{\frac{1}{m_c} \sum_{k=1}^{m_c} (\mu_c - d_k)^2} \tag{2}$$

where m_c is the number of consonantal intervals, μ_c is the mean duration of consonantal intervals, and d_k is the duration of the kth consonantal interval. The standard deviation is scaled by a factor of 100. ΔV is measured identically over the vocalic intervals of the utterance. Like $\%V$, ΔV and ΔC are global in nature, and are not sensitive to the order in which the vocalic or consonantal intervals occur, only the distribution of their durations.

Ramus et al. assert that these metrics capture properties that are psychologically plausible for infants to recognize, and could allow them to discover the principles of rhythm and syllable structure in their native language. This knowledge, in turn, provides cues for the segmentation of the speech signal into stress feet, syllables, or moras, and for the discovery of word boundaries.

Support for these metrics was found in a production and a classification task. In the former, the authors measured $\%V$ and ΔC in sentences elicited from speakers of eight languages (English, Dutch, Polish, Spanish, French, Italian, Catalan and Japanese) and showed that the scores from these two measures group together English, Dutch and Polish on the one hand, and Spanish, French, Italian and Catalan on the other, with Japanese being separated from both groups. The groupings support the idea of distinct rhythm classes in accordance with the consensus classifications for most of these languages. Further, they provide support for assigning Polish and Catalan, previously said to have mixed rhythm, to the stress-timed and syllable-timed groups respectively. Ramus et al. also used a logistic regression on $\%V$ to attempt to discriminate between pairs of languages, finding that discrimination was best between languages from different rhythm classes – although accuracy was still quite modest, ranging between 60 and 65%. The exception was Japanese, which was discriminated from other languages at rates of 90% or more.

Ling, Grabe, and Nolan (2000) introduced the pairwise variability index or PVI, which makes pairwise comparisons between successive consonantal or vocalic intervals. A rate-normalized variation of this metric was later introduced (Grabe & Low, 2002). In contrast to $\%V$, ΔC and ΔV, the PVI metrics are sensitive to the order in which the intervals occur. Although either metric may be used for either kind of interval, in practice, raw PVI (rPVI) is applied to consonantal intervals, and the rate-normalized version of the metric (nPVI) is used for vocalic intervals. The assumption is that vocalic intervals are more prone to variation due to speaking rate, a factor that the authors wanted to eliminate from their measure (Grabe & Low, 2002).

rPVI is the average difference in duration between successive intervals, calculated as follows:

$$\text{rPVI} = (\sum_{k=1}^{m-1} |d_k - d_{k+1}|)/(m-1) \tag{3}$$

where m is the number of intervals, and d_k is the duration of the kth interval. nPVI is similar, except that the difference in duration between successive intervals is expressed as a fraction of the mean duration of those intervals, in an attempt to normalize for tempo:

$$\text{nPVI} = 100 \cdot (\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{\frac{d_k + d_{k+1}}{2}})/(m-1) \tag{4}$$

Dellwo (2006) were also concerned about the correlation between speech rate and rhythm metrics, especially ΔC. They therefore devised a variation coefficient for ΔC, expressing the standard deviation of consonantal interval durations as a percentage of mean consonantal duration, as shown in (5):

$$\text{VarcoC} = 100 \cdot \frac{\Delta C}{\mu_c} \tag{5}$$

where μ_c is the mean consonantal interval duration in the sample. An analogous variation coefficient for ΔV, VarcoV, can be computed in the same fashion. Dellwo claims that VarcoC is superior to ΔC in differentiating French from English and German when speech samples are taken across a range of speaking rates.

Other researchers have proposed additional rhythm metrics, such as the "rhythmogram" algorithm from the auditory prominence hypothesis (Lee & Todd, 2004). $\%V$-ΔC and the PVIs remain by far the most prominent.

2 Weaknesses in the case for the rhythm typology

Infant discrimination studies have provided some of the strongest evidence in favor of the rhythm typology, but these have relied heavily on data from the prototypical languages of each rhythm class: typically Germanic languages such as English and Dutch are used to represent stress-timing, a Romance language such as Spanish or Italian represents syllable-timing, and Japanese represents mora-timing. Demonstrating that infants can differentiate low-pass filtered speech from these limited language groups is an interesting finding but doesn't speak directly to the universality of the rhythm classes.

Further, there is evidence that humans may be using factors other than rhythm when distinguishing between degraded speech signals such as those used in the infant studies. Rodriguez and Arvaniti (2012, in prep.) tested the ability of listeners to differentiate English sentences from Danish, Spanish, Greek, Korean, and Polish sentences, where consonantal intervals had been replaced with a synthesized [*s*] sound, and vocalic intervals replaced with a synthesized [*a*]. They further altered the stimuli in some trials by homogenizing the tempo and/or fundamental frequency (F0) of the sentences. For all languages, discrimination from English was quite poor with tempo and and F0 removed, often not significantly different from chance. Greek and Polish, whose tempos differed the most from that of English, showed markedly better differentiability when tempo information was retained. Greek and Korean were easier to distinguish from English when F0 information was present, compared with the condition with no F0 or tempo information.

These results indicate that both tempo and F0 are useful cues that humans may exploit when judging the similarity of degraded speech signals. This weakens the impact of several of the major empirical results said to support the rhythm typology, including Ramus, Dupoux, and Mehler (2003), which used the same experimental setup as Rodriguez and Arvaniti (2012, in prep.) but without controlling for tempo[2]; and Nazzi et al. (1998), whose infant study used low-pass filtering at 400-Hz, preserving both tempo and F0. It remains to be shown that significant levels of discrimination between languages can be obtained on the basis of pure rhythm, without the inclusion of tempo and F0 information.

The modern rhythm metrics have considerably bolstered the reputation of the rhythm class hypothesis by attempting to re-establish the empirical basis of the typology which was lost when isochrony was debunked. The metric's support for a universal, categorical rhythm typology is far from airtight, however. The case for the metrics and the typology they model is incomplete in several important ways.

Fundamentally, the justification for the metrics themselves often relies on the rhythm typology they are designed to support, a troubling circularity. No external criteria guarantee that the proposed metrics actually measure the impressionistic phenomenon that humans perceive as rhythm. Instead, success in reproducing the traditional rhythm classifications for even a few languages is taken as validation that the metrics are in fact measuring speech rhythm. At the same time, the success of the metrics is seen as support for the rhythm typology itself. The metrics are perceived as truly measuring rhythm because they agree with the rhythm typology, and the typology is empirically confirmed because the rhythm metrics validate it. This is particularly unconvincing because the metrics were specifically constructed with the rhythm hypothesis in mind. Confirmation bias may have operated to ensure that any metrics that were tested and found incompatible were abandoned, under the presumption that they were poor measurements of rhythm.

A certain degree of skepticism is warranted when evidence from a handful of major European languages, largely from the Germanic and Romance families, is cited to assert the existence of a universal typology. And indeed, even the most popular metrics fail to generalize reliably to novel languages beyond the well-studied prototypical exemplars (Arvaniti, 2009). In a study by Grabe & Low of 18 languages, none of the nine unclassified languages examined could be unambiguously assigned to an existing rhythm class. Furthermore,

[2] Some attempt was made to control for tempo in Polish, which was found to be an outlier, but tempo was not controlled for at all in the other languages.

in several cases, class affiliation depended on the choice of metrics used for classification: for example, Thai was classified as stress-timed by PVIs, but as syllable-timed by *%V-ΔC* (Grabe & Low, 2002).

Assuming concerns about metric validity and typology generalization were to be successfully addressed, current work would still fall short of establishing the stronger claims of the rhythm class hypothesis. The metrics studies do little to confirm, and often tend to disconfirm, notions such as the categorical and exclusive nature of rhythm class membership, rhythm as a first-class primitive of speech (not merely an emergent property derived from other factors such as syllable structure and tempo), and the centrality of rhythm to production and comprehension.

Computational modeling has produced some evidence that tends to weaken the metric's support for a categorical rhythm typology. A study using a Bayesian Quadratic Forest Classifier on automatically segmented speech samples from English, Greek, Russian, French and Taiwanese Mandarin found that "[c]lassifiers based on single measure could not distinguish between languages traditionally assigned to different rhythm classes (e.g. English and French) any better than between languages from the same rhythm class (e.g. English and Russian)" (Loukina, Kochanski, Shih, Keane, & Watson, 2009). Rhythm through the lens of the metrics appears rather more like a continuum than a few discrete categories, although it has been proposed that this may be due to there being a larger number of rhythm classes than originally supposed, perhaps five or more (Ramus et al., 1999).

These inconsistencies in the rhythm metrics are reflected in a series of production and perception studies of rhythmically prototypical and non-prototypical languages. In her production study, Arvaniti (in press-b) elicited data from rhythmically prototypical languages – English and German, both considered stress-timed, and Italian, and Spanish, both considered syllable-timed – and non-prototypical Korean and Greek, neither of which has a consensus classification for rhythm type. Varied speech samples were collected from the six languages and each sentence was coded with values for three sets of rhythm metrics: *%V* and *ΔC*, rPVI and nPVI, and VarcoV and VarcoC. Arvaniti found that between-language variation in the metric values were often smaller than the variation between different speakers or different elicitation methods within a single language (Arvaniti, in press-b). These results suggest that within-language variability in timing patterns may be larger than suspected and that metrics are sensitive to it. If the timing metrics don't clearly suggest where the rhythm class boundaries lie, with speakers from the same language appearing less similar to each other than to speakers from other languages, it undermines the argument that the metrics provide proof that a rhythm typology does in fact exist.

This production study was followed by perception experiments in which listeners were asked to rate degraded sentences from the same six languages for similarity to a computer-generated trochee beat (Arvaniti, in press-a), predicting that the stress-timed languages, English and German, would be rated as more similar to the trochee beat than the others (since their foot-based rhythm, which implies an alternation of stressed and unstressed syllables, is closer to a trochee than a cadence, the assumed rhythm of syllable-timed languages). When low-pass filtering was used to degrade the stimuli, all languages were rated in the middle of a 7-point scale and were statistically indistinguishable, except for English which was rated less trochee-like than the other languages. Different results were obtained when the stimuli were replaced by flat *sasasa*. This is the same manipulation used in Rodriguez and Arvaniti (2012, in prep.), and in Ramus et al. (2003). In all three, all consonantal intervals were replaced by [s] and all vocalic intervals by [a], but in Arvaniti (in press-a) the fundamental frequency was flattened in all conditions. In this experiment, English and German were rated more trochee-like than Korean, Greek and Italian, but Spanish was rated as trochee-like as English and German. Overall, these results do not support the view that languages fall into distinct rhythm classes in production or that they are perceived as if they do.

3 Data and methodology

In this study we applied a graph-based clustering engine and a Naive Bayes classifier to the production data from Arvaniti (in press-b). We aimed to test whether the rhythm metric values are consistent with the conception of rhythm as a fundamental, universal, categorical property of human language. We expect that if the rhythm typology is valid and the metrics are measuring rhythm, the clusters identified should correspond to the accepted rhythm classes, and the classifier should be able to consistently distinguish between samples from different classes. Incompatibility between between the metrics and the typology suggests problems with the rhythm hypothesis, the metrics, or both.

To test whether the metrics support the typology, we focused on four central questions:

First, does unsupervised clustering of rhythm metric data produce groupings consistent with the putative rhythm classes? If so, this would be consistent with the typology, and with reports that infants can discriminate between languages on the basis of timing information such as the metrics are based on.

Second, are supervised machine learning models capable of using rhythm metrics to reliably discriminate between data from languages of different rhythm classes? Such a result would support previous research that shows statistical language classification, such as Ramus et al. (1999). The degree of accuracy in this task also forms an important point of comparison for the experiments that follow.

Third, which sets of metrics are most effective in the rhythm classification task? Better-performing metrics might be assumed to be capturing information that is more relevant to the rhythm class distinctions, but there may be confounding factors. For instance, the Varco metrics and nPVI are normalized to remove the effect of speaking rate, while rPVI, $\%V$ and ΔC are not. To what degree do the various metrics correlate with tempo, and is tempo information helpful or harmful in the classification task? Because rhythm is considered to be independent of speech rate, metrics that are normalized for tempo are expected to be as useful in classifying languages for rhythm as their unnormalized counterparts. If tempo is crucial for accurate classification, perhaps the typology reflects speaking rate rather than rhythm.

Fourth, do the models generalize to consistently cluster and classify the data from Greek and Korean, which have no consensus rhythm classification? If so, this would bolster the idea of categorical rhythm classes as a linguistic universal, not merely a property of certain specific languages. If not, this suggests that the typology is not categorical, or it is not universal, or that the metrics do not sufficiently measure rhythm.

3.1 Data

The data used in this study came from previous work by Arvaniti (in press-b). To create the dataset, productions were elicited in six different languages: English and German, which are traditionally classified as stress-timed; Italian and Spanish, which are traditionally classified as syllable-timed; and Greek and Korean, whose classifications are unclear (Arvaniti, 2009). Here we will adopt the designations stress-timed, syllable-timed and unclassified to refer to these groups of languages, without endorsement of the rhythm class hypothesis.

Spontaneous and scripted speech was elicited from eight native speakers of each of these six languages. The speakers performed the following three tasks:

1. They produced approximately a minute of spontaneous speech on a supplied topic. The experimenters worked with the subjects to ensure that each had a topic that they could speak about easily and fluidly. Topics used included parking on campus, airport security, daily life in the city, and the description of a scenes from a comic strip.
2. They read Aesop's story of the North Wind and the Sun. The story, five sentences long, was supplied to them in translation in their native language.

3. They read 15 isolated sentences, between fifteen and twenty-five syllables in length. For each language, the set of sentences was divided into three sub-sets of five sentences each:
   (a) Stress-timed sentences: These sentences were constructed by the experimenters to be maximally compatible with stress-timing. This was achieved by selecting words with a great deal of variability in syllabic structure, including consonant clusters, geminates, diphthongs and other features as appropriate for the language. Because stress-timing is purported to involve a greater variation in the duration of consonantal and vocalic intervals, the greater complexity and variety of forms in these sentences favor this class.
   (b) Syllable-timed sentences: Chosen using the opposite criteria as the stress-timed sentences, this group exhibited simple syllable structure, containing as few consonant clusters, diphthongs and geminates as possible. The simpler and less varied syllables in this set favor the lower variation in consonantal and vocalic interval duration ascribed to syllable-timing.
   (c) Uncontrolled: These sentences were randomly selected from existing (mostly literary) works in the language. The criteria for inclusion were only length and some degree of meaningfulness when taken out of context, and they were therefore uncontrolled for timing.

The collected utterances were manually annotated for the duration of each consonantal and vocalic interval, and values for $\%V$, ΔC, nPVI, rPVI, VarcoV and VarcoC were calculated based on these values. These metric values were the input to the present study. For complete details of materials and elicitation and annotation procedures, see Arvaniti (in press-b).

In our Naive Bayes classifications, we used only the data from the spontaneous corpus and the uncontrolled sentences, avoiding the artificially manipulated sentences. The within-language variation that was intended to be introduced by these manipulations could artificially bias the classifier. The entire corpus was used in the unsupervised clustering experiments, where language and rhythm class labels were not given to the clustering algorithm and hence could not influence cluster membership. Although the manipulated sentences doubtless had some effect on the overall clustering solution, they were not generally outliers and were assigned to the cluster into which they naturally fell, allowing us to examine whether the manipulations had the intended effect.

4 Methods

We performed two separate series of experiments. In the first, we performed unsupervised clustering, asking the clustering algorithm to group the data without knowledge of the language or rhythm class labels. The generated clusters were then analyzed to see how closely they corresponded with the language and rhythm class groups from which the data was drawn. In the second series of experiments, we performed supervised learning using a Naive Bayes classifier. Our aim was first to assess the ability of the models to discriminate between different languages and different rhythm classes on the basis of the timing data encoded by the metrics; second, to evaluate the relative contributions of each of the three metrics sets to the classifier's accuracy; and third, to test the ability of the model to predict rhythm class for the unclassified languages Greek and Korean.

4.1 Clustering

Clustering is an unsupervised learning method which seeks to identify and group together regions of high similarity in a data set. The specific method we used is a minimun cut graph partitioning algorithm, as implemented in the CLUTO software package (Karypis, 2003). This clustering procedure begins by creating a graph node for each data instance (each sentence, in our case). For each node, the 40 nearest neighbors

are found, in terms of Euclidean distance. An edge is drawn between each pair of nodes that are symmetrically amongst each other's nearest neighbors. When all the appropriate edges have been drawn, the graph is separated into two components by drawing a line (which need not be straight) through as few edges as possible while creating two good clusters. "Good" in this context means maximizing internal similarity while minimizing external similarity.

Clustering has some advantages over other modeling approaches for our purposes. First of all, unlike the supervised methods of regression or Naive Bayes, it requires no prior information about the rhythm class membership of the sentences it clusters. It is a way of checking whether the metric data naturally model the rhythm typology, without having to rely on labels that already assume that the rhythm typology exists. This can be seen as more closely approximating infant discrimination studies such as the one performed by Nazzi et al. (1998), where infants are exposed to low-pass filtered speech in two foreign languages, without any information about which stimuli belong to which language.

Secondly, clustering is capable of learning models where the underlying classes in the data are not linearly separable. This allows for considerably more flexibility than linear models such as Naive Bayes. This is a major advantage if the relationship between the metrics and the rhythm classes is not linear.

We used the CLUTO clustering package (Karypis, 2003) for all clustering runs. The choice of the Euclidean distance-based min-cut partitioning procedure was motivated by the CLUTO author's report of the superior performance of this technique on low-dimensionality data sets like ours.

4.2 Naive Bayes classification

Naive Bayes is a supervised learning algorithm that learns a classification model from a collection of labeled data instances, and then can predict the labels of future unseen instances from the distribution. Although independence of variables is technically a requirement for Naive Bayes models, they are quite robust even when the data used contains dependent variables, and have proven to be accurate in classifications in a wide range of scenarios (Witten & Frank, 2005).

The Naive Bayes model is built from a set of labeled data instances, where each instance has observed values for a number of variables, and a label specifying what class it belongs to. In our case, the instances represent sentences and their associated values are those of the rhythm metrics. For certain experiments, the class labels correspond to a single language (e.g. Italian vs. German, or Spanish vs. Italian), and in some cases, the class labels are the rhythm classes themselves, syllable-timed and stress-timed (i.e. when classifying English and German vs. Spanish and Italian).

A Naive Bayes classifier takes advantage of the central insight captured by Bayes' theorem, which is that it is possible to make predictions about the causal factors behind an event, based on the outcome of the event and some knowledge of the likelihood of the causal factors to produce that outcome. In our case, the hypothesis under investigation is that rhythm class is an important causal factor influencing the metric values recorded for the sentences. If that hypothesis is correct, it should be possible to reliably predict the rhythm class for sentences based on their metric values, through the use of Bayes' theorem. The probability of a sentence belonging to a certain class, given its values for the metrics, can be estimated if we can estimate the probability of the metrics taking on those values for a sentence in that class, the prior probability of a sentence belonging to that class, and the prior probability of the metrics taking on the observed values. This is shown by the formulation of Bayes' theorem in (6).

$$P(class|metrics) = \frac{P(metrics|class)P(class)}{P(metrics)} \tag{6}$$

Because we are only interested in finding out which class is most probable for a given sentence, we only care about the relative magnitudes of the probabilities for each class, and we can ignore any part of the equa-

tion that is not dependent on the class. The prior probability of the metrics in the denominator, $P(metrics)$, is the same for each class, so it can be ignored.

In our experiments, we have only used balanced training data when building our models, with equal numbers of sentences from each language or rhythm class. Therefore, the prior probability of a sentence belonging to a particular class is always the same, 0.5. Thus we can ignore $P(class)$ as well, leaving the simplified formula in (7). In cases where an estimate of prior class probability is needed, it is common to use the maximum likelihood estimate, which is simply the relative frequency of that class in the training data.

$$P(class|metrics) \propto P(metrics|class) \tag{7}$$

We only need to estimate the joint probability of the observed metric data, given the class, in order to compute estimates for the relative probability of the sentence belonging to each class. The "naive" assumption of the Naive Bayes classifier simplifies this calculation by allowing us to assume that all metrics are independently distributed (although we are aware they are not). This allows us to estimate the joint probability of the observed metric data as the product of the individual metric probability estimates. (8) shows the general form of the joint probability model for the posterior probabilities of the observed metrics values (M_1 - M_n) given the class (C), which must be used if considering the metrics as dependent variables; (9) shows the simplified version allowed by the naive assumption of independence.

$$P(M_1|C)P(M_2|M_1,C)...P(M_n|M_1,M_2,...,M_{n-1},C) \tag{8}$$

$$P(M_1|C)P(M_2|C)...P(M_n|C) \tag{9}$$

(9) gives us a convenient method for finding an estimate of the joint probability of the metric values, so the remaining task is to estimate the probabilities of the individual metrics. Estimation of the probability of continuously valued variables can be done through discretization and relative frequency estimation, assuming a Gaussian distribution and estimating its parameters with maximum likelihood estimation, or using a non-parametric method such as LOESS, which is the method used in our models. LOESS estimates the probability of an unseen data point using a locally weighted regression over the probabilities of the data points in a window around the unseen point, weighting the nearest points so that they are more important to the prediction. In this way it is less sensitive to outliers in the distribution.

The decision procedure for assigning a class label to a sentence is then to find the class label that maximizes the estimated probability for the observed metric values for that sentence, as represented in (10).

$$\underset{class}{\operatorname{argmax}}\, P(M_1|class)P(M_2|class)...P(M_n|class) \tag{10}$$

In our case, all of our classifications are binary, so it's simply a matter of trying both classes and choosing the class with the higher estimated posterior probability.

The Naive Bayes classifier used in this study is from the Orange machine learning environment 2.0 Beta release (Curk et al., 2005). All Naive Bayes classification accuracies presented in this paper were computed using leave-one-out cross validation, except as noted.

5 Results

5.1 Clustering

In these experiments we compared the groups formed by the clustering algorithm with the groups predicted by the rhythm class typology. If the sentences from each rhythm class cluster together, this would support the

view that languages inherently belong to a particular rhythm class, and that the metrics capture this distinction. It would also be compatible with experiments that show that infants are able to distinguish between languages from different rhythm classes, but not between languages from the same rhythm class (Nazzi et al., 1998; Nazzi & Ramus, 2003).

We ran experiments using each of the three metric sets (*%V-ΔC*, PVIs, and Varcos) separately, and with all metrics simultaneously. The results for the two cluster runs are given in Table 1.

For *%V* and *ΔC*, cluster two includes the vast majority of the Italian and Spanish sentences, the majority of English syllable-timed sentences, and the lion's share of Greek and Korean sentences as well. Clearly cluster two is dominated by the syllable-time productions. Greek and Korean appear to pattern much more closely with the syllable-timed languages in the metric space defined by *%V* and *ΔC*.

The situation is nearly identical for the PVI metrics, with most of English and German preferentially placed in cluster one, and almost everything else in cluster two.

The Varco metrics show no clear pattern except that cluster two is somewhat larger, with majorities of all languages appearing there. The clustering solution does nothing to distinguish syllable-timed from stress-timed languages, or to predict a rhythm class for Greek or Korean.

Using all metrics, the clusters are quite similar to those seen for *%V-ΔC* and the PVI metrics, with cluster two being much larger and dominated by syllable-timed languages, with Greek and Korean heavily represented as well. Overall, *%V-ΔC* and the PVIs appear to do a good job of grouping the syllable-timed and unclassified sentences together, with most stress-timed sentences in a different cluster.

Effects of elicitation type can be readily seen in the clustering solutions. For *%V-ΔC*, cluster one contains almost no sentences from the syllable-timed languages, so although it only attracts slight majorities of English and German, it is clearly dominated by them. Strikingly, it does manage to attract 87.5% of English sentences that are manipulated to be stress-timed, and 95% of such German sentences, the largest shares it gets from any sentence groups in any language. Conversely, syllable-timed sentences from English and German are the most likely to fall into cluster two: 82.5% of the English and 47.5% of the German syllable-timed sentences are found there. These manipulated sentences are clearly having the predicted effect, shifting the metric values towards the rhythm class they were targeted for. For the syllable-timed languages, stress-timed sentences are shifted toward cluster one in Spanish but not in Italian, and the effect of the syllable-timed sentences cannot be seen, since almost all the sentences from all sentence types are in cluster two.

6 Naive Bayes Results

6.1 Naive Bayes Experiment 1: Classification by language and rhythm class

In this experiment we asked a basic question: can speech samples be accurately classified by language and rhythm class on the basis of the information provided by all rhythm metrics together? The expectation is that if the rhythm typology is correct, and the metrics capture the relevant properties of rhythm, it should be possible to predict rhythm class membership from the metric values. We trained separate models to distinguish between each pair of languages. Additionally, we grouped the languages together into stress-timed, syllable-timed and unclassified languages, and trained a separate model to distinguish between each pairing of these groups. Models were generated separately for the spontaneous and the uncontrolled sentences.

Our results indicate that a Naive Bayes learner is capable of achieving fairly high cross-validated accuracy in distinguishing between the prototypical languages from opposing rhythm classes. Discrimination is best, at around 80% accuracy, between languages from different rhythm classes. Similarly, when the data were grouped by rhythm class, discrimination was at 81.9% accuracy between the stress-timed and syllable-time languages. The models were less accurate in discriminating between languages from the same rhythm class, with Italian/Spanish accuracy falling as low as even chance on the spontaneous sentences.

Table 1. Clustering all sentences for all elicitation methods into two clusters. Each double column represents a separate clustering run, using the metrics indicated. Cell values are the percentage of instances for that language and data type that were grouped into that cluster. Clusters containing 70% or more of instances for a given data type are bolded.

		%V-ΔC		PVI		Varco		All Metrics	
Language	Elicitation	C1	C2	C1	C2	C1	C2	C1	C2
English	Spontaneous	64.3	35.7	68.8	31.2	61.6	38.4	66.1	33.0
	Northwind	50.0	50.0	44.6	55.4	21.4	**78.6**	41.1	58.9
	Uncontrolled	45.0	55.0	37.5	62.5	12.5	**87.5**	42.5	57.5
	Stress	**87.5**	12.5	**90.0**	10.0	15.0	**85.0**	**82.5**	17.5
	Syllable	17.5	**82.5**	35.0	65.0	17.5	**82.5**	22.5	**77.5**
German	Spontaneous	**76.0**	24.0	64.6	35.4	39.6	60.4	67.7	32.3
	Northwind	**73.4**	26.6	53.1	46.9	26.6	**73.4**	53.1	46.9
	Uncontrolled	55.0	45.0	35.0	65.0	30.0	**70.0**	40.0	60.0
	Stress	**95.0**	5.0	**85.0**	15.0	12.5	**87.5**	**72.5**	27.5
	Syllable	52.5	47.5	47.5	52.5	17.5	**82.5**	35.0	65.0
Italian	Spontaneous	9.2	**90.8**	11.5	**88.5**	60.9	39.1	10.3	**89.7**
	Northwind	2.1	**97.9**	6.2	**93.8**	39.6	60.4	6.2	**93.8**
	Uncontrolled	2.5	**97.5**	10.0	**90.0**	22.5	**77.5**	10.0	**90.0**
	Stress	2.5	**97.5**	10.0	**90.0**	7.5	**92.5**	5.0	**95.0**
	Syllable	2.5	**97.5**	15.0	**85.0**	12.5	**87.5**	7.5	**92.5**
Spanish	Spontaneous	15.0	**85.0**	28.8	**71.2**	63.8	36.2	22.5	**77.5**
	Northwind	8.3	**91.7**	20.8	**79.2**	14.6	**85.4**	6.2	**93.8**
	Uncontrolled	10.0	**90.0**	22.5	**77.5**	40.0	60.0	22.5	**77.5**
	Stress	41.5	58.5	36.6	63.4	2.4	**97.6**	29.3	**70.7**
	Syllable	0.0	**100.0**	7.5	**92.5**	10.0	**90.0**	0.0	**100.0**
Greek	Spontaneous	17.9	**82.1**	17.9	**82.1**	60.3	39.7	16.7	**83.3**
	Northwind	3.3	**96.7**	8.3	**91.7**	46.7	53.3	10.0	**90.0**
	Uncontrolled	0.0	**100.0**	5.0	**95.0**	57.5	42.5	0.0	**100.0**
	Stress	7.5	**92.5**	10.0	**90.0**	40.0	60.0	7.5	**92.5**
	Syllable	0.0	**100.0**	0.0	**100.0**	15.0	**85.0**	0.0	**100.0**
Korean	Spontaneous	28.7	**71.3**	41.7	58.3	**75.0**	25.0	33.3	65.7
	Northwind	15.6	**84.4**	18.8	**81.2**	31.2	68.8	20.3	**79.7**
	Uncontrolled	15.0	**85.0**	22.5	**77.5**	12.5	**87.5**	22.5	**77.5**
	Stress	25.0	**75.0**	27.5	**72.5**	20.0	**80.0**	35.0	65.0
	Syllable	35.9	64.1	35.9	64.1	23.1	**76.9**	38.5	61.5

However, accuracies were still significantly better than chance for English and German, at 75% for the uncontrolled sentences. This was only marginally worse than English/Spanish discrimination on the uncontrolled sentences, and actually better than English/Spanish discrimination on the spontaneous sentences, which was performed at 72.2% accuracy.

Discrimination between the two unclassified languages Greek and Korean had the highest overall classification accuracy, at 85%, suggesting that they are not members of the same class. Accuracy was worst overall between Korean and Spanish, and Korean and Italian, suggesting that Korean metrics pattern with those of the syllable-timed languages.

Table 2. Classification accuracies of Naive Bayes for pairwise comparison models, using data for all metrics from the uncontrolled and spontaneous sentences

	Stress-Timed		Unclassified		Syllable-Timed	
	English	German	Greek	Korean	Italian	Spanish
English	–	75.0%	83.8%	73.8%	82.5%	76.3%
German	75.0%	–	82.5%	82.5%	78.8%	83.8%
Greek	83.8%	82.5%	–	85.0%	81.3%	73.8%
Korean	73.8%	82.5%	85.0%	–	56.3%	55.0%
Italian	82.5%	78.8%	81.3%	56.3%	–	61.3%
Spanish	76.3%	83.8%	73.8%	55.0%	61.3%	–
English	–	70.3%	75.6%	65.7%	81.0%	72.2%
German	70.3%	–	79.5%	77.0%	85.0%	79.6%
Greek	75.6%	79.5%	–	59.6%	63.5%	56.4%
Korean	65.7%	77.0%	59.6%	–	63.2%	54.9%
Italian	81.0%	85.0%	63.5%	63.2%	–	50%
Spanish	72.2%	79.6%	56.4%	54.9%	50%	–

Table 3. Classification accuracies of Naive Bayes for rhythm class, using data for all metrics from the uncontrolled sentences

	Stress-timed	Unclassified	Syllable-time
Stress-time	–	71.9%	81.9%
Unclassified	71.9%	–	61.3%
Syllable-time	81.9%	61.3%	–
Stress-time	–	71.5%	80.4%
Unclassified	71.5%	–	57.7%
Syllable-time	80.4%	57.7%	–

6.2 Naive Bayes Experiment 2: Comparison of metrics in prototypical languages

In Experiment 1, our models were trained using all three sets of metrics (*%V* and *ΔC*, PVIs and Varcos) concurrently. In Experiment 2, we examined the performance of each metric set separately, to expose their

individual contributions to classifier accuracy. We compared their performance on the most accurate task from Experiment 1, classifying pairs of languages from different rhythm classes. Four pairwise models were trained, one for each possible pairing of one stress-timed language (English or German) and one syllable-timed language (Italian or Spanish). An additional model was trained to distinguish both stress-timed languages from both syllable-timed languages, for a total of five model types. Each of these model types was trained on each of the three metric sets separately, plus all metrics combined, for a total of 15 models. The entire set of experiments was run twice, once on the uncontrolled sentence data and once on the spontaneous data, for a grand total of 30 models. The results are presented in Table 4.

Table 4. Classification accuracies in discriminating between stress-timed and syllable-timed languages using each set of metrics separately. The most accurate metric type for each comparison is bolded. The accuracy obtained when using all metrics together, from Experiment 1, is given for reference in the last column.

Elicitation	Task	*%V-ΔC*	PVI	Varco	All Metrics
Uncontrolled	Eng vs. Ita	76.2%	**77.8%**	47.5%	82.5%
	Eng vs. Spa	**68.6%**	66.3%	61.3%	76.3%
	Ger vs. Ita	**81.3%**	70.0%	58.8%	78.8%
	Ger vs. Spa	**77.5%**	65.0%	56.3%	83.8%
	Ger/Eng vs. Ita/Spa	**75.6%**	70.6%	46.9%	81.9%
Spontaneous	Eng vs. Ita	**77.4%**	75.9%	56.3%	74.6%
	Eng vs. Spa	**71.0%**	70.0%	62.2%	71.6%
	Ger vs. Ita	**85.3%**	74.3%	63.4%	84.7%
	Ger vs. Spa	**79.1%**	69.5%	63.3%	80.8%
	Ger/Eng vs. Ita/Spa	**78.5%**	71.8%	62.7%	80.3%

Table 4 shows that *%V-ΔC* is nearly always the most predictive set of metrics, and using those metrics alone tends to yield classification accuracies within a few percentage points of those afforded by using all metrics together. PVIs follow, generally around 4-6% behind, and the Varco-based classifiers are far behind at the bottom of the pack.

In three instances, the accuracies obtained from the *%V-ΔC* data alone actually slightly exceeded those obtained from classifiers built using all three metric sets, as in the case of Spontaneous: English vs. Italian in Table 4. The Naive Bayes classifier is capable of being misled by extra data that is either irrelevant to the classification task, or duplicates information already present in the data set. The inclusion of multiple highly correlated variables in the model with all metrics is a rather strong violation of the (already naive) assumption of independence, and it may be that these dependent variables are forcing a less optimal solution. Another possibility is that the omission of the less-than-helpful Varco data allows the classifier to converge on a more accurate model.

6.3 Experiment 3: Classifying the unclassified languages

In Experiment 3, we looked at the performance of the Naive Bayes classifier in assigning Greek and Korean, which are of undetermined rhythm class, to either stress- or syllable-timing. We used only the classification model that achieved the best overall accuracy in discriminating between stress-timed and syllable-timed data in Experiments 1 and 2, which was the model that used all three metric sets.

Separate models were trained to discriminate between each possible combination of one stress-timed language (English or German) and one syllable-timed language (Italian or Spanish), yielding four separate models. One additional model was trained to distinguish all stress-timed data (English and German) from all syllable-timed data (Italian and Spanish). Each model was then asked to predict the class of each sentence in the Greek and Korean data, and the percentage of data instances assigned to each class was recorded.

Table 5. Unclassified language classifications for models trained on data from stress-timed vs. syllable-timed languages, from the uncontrolled and spontaneous sentences. The majority class is bolded. Here leave-one-out cross-validation is only applicable to the "Accuracy" column; other columns show rates of classification where accuracy cannot be computed, and no cross-validation is possible.

Model Training Data			Greek		Korean	
Stress	Syllable	Accuracy	Stress	Syllable	Stress	Syllable
Eng	Ita	82.5%	37.5%	**62.5**%	42.5%	**57.5**%
Eng	Spa	76.3%	30.0%	**70.0**%	50.0%	50.0%
Ger	Ita	78.8%	**60.0**%	40.0%	25.0%	**75.0**%
Ger	Spa	83.8%	25.0%	**75.0**%	32.5%	**67.5**%
Eng+Ger	Ita+Spa	81.9%	32.4%	**67.6**%	37.5%	**62.5**%
Eng	Ita	81.0%	23.0%	**76.9**%	35.1%	**64.8**%
Eng	Spa	72.2%	21.8%	**78.2**%	37.0%	**63.0**%
Ger	Ita	85.0%	19.2%	**80.8**%	31.5%	**68.5**%
Ger	Spa	79.6%	19.2%	**80.8**%	28.7%	**71.3**%
Eng+Ger	Ita+Spa	80.4%	21.8%	**78.2**%	35.2%	**64.8**%

The Greek and Korean data are generally classified with the syllable-timed languages. The trend is weak and inconsistent in the uncontrolled data, and stronger and more consistent for the spontaneous sentences. The trends are stronger for Greek than for Korean.

For the uncontrolled data, the class and strength of classification varied depending on which specific languages were used to train the model. For instance, the models trained with Spanish as the syllable-timed language classified Greek as moderately syllable-timed, while those that used Italian classified it as weakly syllable-timed or weakly stress-timed. The model that achieved 82.5% accuracy in distinguishing English from Italian was nearly evenly split on the Korean data, classing it with English 42.5% of the time and with Italian 57.5% percent of the time. The effect of specific language was not as pronounced in the spontaneous speech data, but Korean still shows a greater tendency to be classified as syllable-timed when the stress-timed member of the training set is German. This is consistent with the results shown in Table 2, where Korean is more easily distinguished from German than from English.

Overall, the metrics show some ability to generalize and predict rhythm class membership for new languages, but the effect is weak and inconsistent for the uncontrolled data. If the metrics do accurately reflect rhythm, the prediction would be that Greek has a rhythm most like Spanish and least like German, while Korean's rhythm is somewhat more syllable-timed than stress-timed. The major effect of elicitation type is unexplained and puzzling.

7 Discussion

7.1 Clustering

The simultaneous two-way clustering of all sentences for all elicitation types showed that *%V-ΔC* and the PVI metrics tended to group English and German together in one cluster, and the syllable-timed and unclassified languages in another. A more nuanced picture emerges, however, if we compare smaller slices of the data in finer detail. We observe that although cluster two tends to include almost all instances of Italian, Spanish, Greek and Korean, it also includes half or more of the English and German sentences for some elicitations. Figures 1 and 2 demonstrate why this would be the case: German values for *ΔC* actually fully bound the Spanish values for this metric, but only German sentences are found in the upper range. Therefore, a reasonable clustering solution will group most Spanish sentences and some German sentences in one lower-range cluster, with only German sentences, those with higher *ΔC* values, in another cluster.

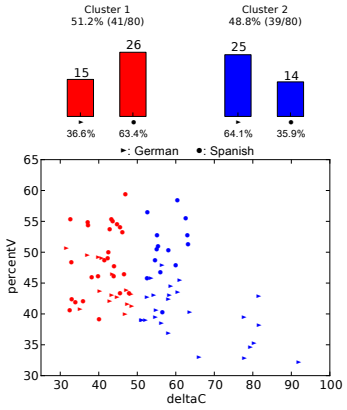

Fig. 1. A clustering solution for *%V-ΔC* metrics on uncontrolled sentences for German and Spanish. Bar charts for each cluster show its composition by language. The count of sentences from the language appears above the bar and the percentage of the cluster constituted by that language appears below.

This kind of overlap between languages is not confined to *ΔC*; in fact, ranges overlap even more dramatically in several cases. See boxplots for all metrics in the appendix in Section 9. Rather than representing a neat separation of rhythm groups, our clustering solutions actually point out a substantial overlap of values, with differing upper and lower bounds for some languages.

Even considering only the mean values, our findings do not support a clean separation of languages along the rhythm typology boundaries. Very frequently, a pair of languages from the same rhythm group were found to have a larger difference in means than some other pair of languages from different rhythm groups, as other researchers have noted (White & Mattys, 2007; Arvaniti, in press-b). For example, using the VarcoC metric on the uncontrolled sentences, the difference between the means of Italian and Spanish was 8.2, while the difference between the Spanish and German means was only 5.7. Table 6 shows languages from the same rhythm group whose means for a given elicitation method and metric differed significantly at $p < 0.05$ in a Tukey HSD, and were farther apart than at least one pair of means from languages from different rhythm groups for the same elicitation and metric.

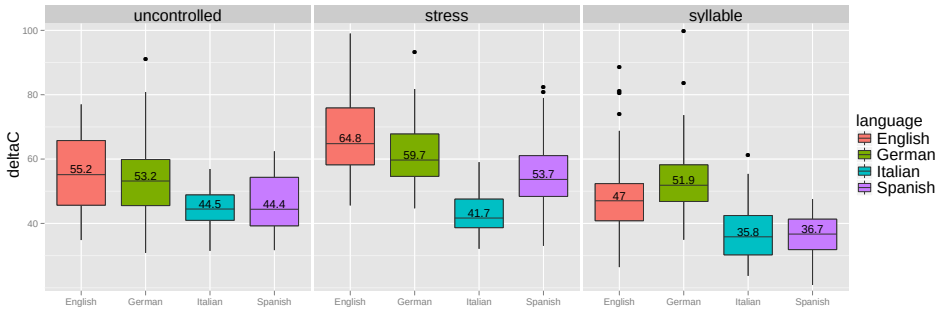

Fig. 2. Boxplot showing ΔC metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

Elicitation	Languages		Metric(s)
Uncontrolled	Italian	Spanish	VarcoC
Stress	German	English	rPVI
	Italian	Spanish	ΔC, rPVI, VarcoC
Syllable	English	German	$\%V$

Table 6. Unexpected differences: the means of these language pairs from the same rhythm group are farther apart than some pairs of languages from different rhythm groups, for the same metric and elicitation method.

Elicitation	Languages		Metric(s)
Uncontrolled	English	Italian	VarcoC, VarcoV
	English	Spanish	$\%V$, rPVI, VarcoV
	German	Italian	rPVI, VarcoC, VarcoV
	German	Spanish	rPVI, VarcoV
Stress	English	Italian	VarcoC, VarcoV
	English	Italian	VarcoV
	English	Spanish	VarcoV
	German	Italian	VarcoV
	German	Spanish	VarcoC, VarcoV
Syllable	English	Italian	$\%V$, VarcoC, VarcoV
	English	Spanish	VarcoC, VarcoV
	German	Italian	VarcoV
	German	Spanish	VarcoV

Table 7. Unexpected similarities: the means of these language pairs from different rhythm groups are closer together than some languages pairs from within a single rhythm group, for the same metric and elicitation.

Conversely, language pairs from different rhythm groups quite often appeared more similar to each other than pairs within the same rhythm group. Table 7 shows language pairs from different rhythm classes whose means are closer to each other than at least one language pair from within the same rhythm class, for a given metric and elicitation type. For example, English is closer to Italian than to German on the stress-timed sentences, as measured by VarcoC and VarcoV.

If we expand the comparison to include distances between languages when elicitation methods differ, the coherence of the rhythm groups deteriorates even further. For example, uncontrolled Italian and German stress-timed sentences were more similar to each other than English syllable-timed and English stress-timed sentences, as measured by rPVI. A full 150 out of 396 pairwise comparisons[3] are anomalous in this way: either the two languages appear to be closer than should be possible for members of different rhythm groups, or they are more distant than should be possible for languages from the same rhythm group.

These findings are not incompatible with a rhythm typology where the categories are broad and different languages are distributed in different ways within their category. Under such a scenario, it's quite possible that certain syllable-timed languages might be found closer to certain stress-timed languages than to other languages in their own class. However, the metrics have been called in to service not merely to characterize the rhythm typology, but also to prove its very existence. An appealing argument on the basis of the metrics could be made if all languages were found to be most highly similar to those in their purported rhythm class, but as it stands, the most that can be said is that the metrics do not rule out certain weaker versions of the rhythm hypothesis.

7.2 Naive Bayes discussion

We were able to construct Naive Bayes models that achieved accuracies greater than 80% in distinguishing between prototypical languages of the two rhythm classes, and between the two classes themselves, on the basis of the rhythm metric data. From this we conclude that the metrics are indeed capturing differences between the prototypically stress-timed and syllable-timed languages, a finding which is generally consistent with the rhythm class hypothesis.

	tempo	percentV	deltaC	rPVI	nPVI	VarcoC	VarcoV
tempo		0.168	**-0.647**	**-0.583**	-0.349	-0.222	-0.208
percentV	0.168		-0.393	-0.339	0.066	0.250	0
deltaC	**-0.647**	-0.393		**0.860**	0.213	0.112	0.753
rPVI	**-0.583**	-0.339	**0.860**		0.236	0.076	**0.611**
nPVI	-0.349	0.066	0.213	0.236		**0.600**	0.099
VarcoC	-0.222	0.250	0.112	0.075	**0.600**		0.136
VarcoV	-0.208	0	**0.753**	**0.611**	0.099	0.135	

Table 8. Correlation matrix for all metrics plus tempo, across all languages and elicitations. Correlations > 0.3, generally considered moderate, are underlined. Correlations > 0.4, considered strong, are **bolded.**

[3] 4 languages combined with 3 elicitation methods yield 12 separate factors. 12 factors assort into 66 unique pairs, examined for each of 6 metrics, generating 396 comparisons. The Tukey HSD method was used to test whether the means at each factor level were significantly different, and only those that were significant at the $p < 0.05$ level were recorded as anomalously distant.

In section 6.3, the models that were trained on the prototypical stress-timed and syllable-timed languages were shown to predict a syllable-timed classification for Greek most of the time, with a weaker tendency toward syllable-time for Korean. The effect of the elicitation method was quite strong, with the spontaneous sentences looking markedly more syllable-timed, especially for Greek.

The predictions were sensitive to the particular languages used to train the model, so that Greek appeared to be weakly stress-timed to the model trained on German and Italian, but was syllable-timed according to the model trained on German and Spanish. This is more evidence of substantial variation between languages within the same hypothetical rhythm class. Such variation does not rule out the existence of a rhythm typology, but it erodes the argument that the rhythm metrics can show us where to draw the lines between rhythm classes.

The metrics are much more effective at replicating the prototypical Germanic-vs-Romance rhythm class distinction than they are at predicting the rhythm classes for novel languages, calling into question the universality of the rhythm typology. This conclusion echoes the results of production studies, such as Grabe and Low (2002) who also showed that many languages are hard to classify for rhythm based on information from metrics.

When we isolated the individual metric sets in Experiment 3, it became clear that the bulk of the predictive power of the models was derived from $\%V$-ΔC, which was generally able to predict rhythm class with an accuracy nearly equal to that obtained by using all metrics concurrently. This finding was echoed in the clustering experiments, where the $\%V$-ΔC-based clusters showed the closest resemblance to the groupings predicted by the rhythm class typology. In contrast, the Varco metrics were both the least effective in discriminating between rhythm classes, rarely breaking 65% accuracy, and the least useful in generating meaningful clusters. The PVIs fell somewhere in the middle in both tasks.

The explanation for the difference in their utility may well be the same as the explanation for why the metrics have any predictive power at all: the metrics model the rhythm typology only to the extent that they correlate with tempo. Neither of the $\%V$-ΔC metrics are designed to normalize for speech rate, one of the PVIs (nPVI) is normalized and one (rPVI) is not, and both Varcos are normalized. Table 8 shows the Pearson's correlation coefficients for each pairing of the metrics and tempo, where tempo is calculated as the number of vocalic intervals per second, discounting any substantial pauses in the utterance. This is a rather rough metric for tempo, and its possible that a more refined metric would show slightly different values. The major trends are clear, however. ΔC, from the most accurate metric set, is strongly correlated with temp. rPVI, from the second most accurate set, slightly less so. The Varcos, essentially useless for prediction or clustering, are only weakly correlated.

Although nPVI is designed to normalize for tempo, it has been objected that nPVI does not in fact effectively do so at a global level, because it normalizes only with respect to the mean duration of two adjacent intervals (Barry, Andreeva, Russo, Dimitrova, & Kostadinova, 2003). It would appear that this objection is warranted, as nPVI is still moderately well correlated with tempo.

The correlations with tempo in the metrics that measure durational variability are not surprising. It is well-known that durational variability diminishes as speaking rate increases. Any segment must be produced with some minimum duration in order to heard, so in faster speech, there is simply less room to vary duration (Dauer, 1983). Given that stress-timed languages like English and German are spoken at slower rates than syllable-timed languages like Spanish and Italian, it is not unlikely that the slower rate of the former group results in their higher durational variability (Arvaniti, 2009; Dellwo & Wagner, 2003).

Evidence for the connection between rhythm and speaking rate comes from a variety of sources. Arvaniti (in press-b) found that normalized metrics, such as Varcos, perform less well than metrics not normalized for speaking rate. Dellwo and Wagner (2003) and Russo and Barry (2008), who calculated metrics separately

for data spoken at different speaking rates, found that the values of metrics decreased as speaking rate increased. Renwick (2012) found that the *%V* metric is quite sensitive to syllable structure, and is strongly correlated with open syllables and negatively correlated with consonant clusters, factors which strongly influence speech rate. Finally, Rodriguez and Arvaniti (2012, in prep.) shows that speaking rate is crucial for language discrimination based on human perception. In a series of experiments, listeners could discriminate much better between English on the one hand and Polish, Spanish, Danish, Greek or Korean on the other when the original speaking rate of the stimuli was retained, especially when the two languages in question had very different speaking rates, like English and Greek. Discrimination was very low or even impossible, however, when speaking rate was normalized across the two languages of the stimuli.

In this study, the better correlated the metrics are with rate of speech, the more accurately they can classify sentences by rhythm group, and the more closely the clusters they produce model the rhythm group typology. Romance languages such as Italian and Spanish are spoken at a higher tempo than Germanic languages such as English and German. Given that the successful metrics are those that are strongly correlated with tempo, and that the classic rhythm typology sees Germanic languages and stress-timed and Romance languages as syllable-timed, it would appear that the success of the metrics is largely or wholly dependent on the extent to which they act as a proxy for tempo. Because rhythm is conceived of as being independent of speech rate, this suggests the metrics are not measuring rhythm at all.

8 Conclusion

We found that the rhythm metrics are able to accurately classify speech data as either stress-timed or syllable-timed when dealing with the prototypical exemplars of each rhythm group. Further, clustering showed that most syllable-timed sentences cluster together, with a majority of the stress-timed sentences in a different cluster. An inconsistent ability to predict rhythm class membership for heretofore unclassified languages was also observed.

These findings are generally compatible with a belief in a weak version of the rhythm typology hypothesis, and with the belief that the metrics are truly quantifying some rhythmic properties of speech. However, it is crucial to note that this simple statement of hypothetical compatibility is as far as the evidence may be taken, and we find nothing to indicate that there must be a rhythm typology of any kind, let along the particular one characterized by universal, categorical rhythm classes of the stress- and syllable-timed variety. Nor do our results drive us toward the conclusion that the metrics are in fact measuring rhythm.

In fact, it is clear that whatever else they may be measuring, some of the metrics are very good measures of speaking rate. The ability of the metrics to classify and cluster in ways compatible with the rhythm typology appears to hold only to the extent to which a given metric set is correlated with tempo. The Varco metrics, the least well correlated, performed the worst, failing completely at every task. The PVIs, strongly correlated with tempo, were much more successful. ΔC, the most strongly correlated, showed the best results in classification and clustering. Given that the stress-timed and syllable-timed languages are easily distinguished by speaking rate, it seems likely that any descriptive power the metrics have is due to the fact that they carry a degraded signal for tempo.

Although the clustering shows that syllable-timed sentences are consistently grouped together, many stress-timed sentences are also included in this cluster. A look at the complete distribution of the metrics reveals the cause: English and German consistently show a wider range of values than Italian or Spanish, while often overlapping the smaller range completely. One cluster forms over the overlapping region, and another smaller cluster contains the English and German sentences in the non-overlapping part of their range.

The argument may be made that the metrics are still valuable because they do show a significant difference in their means between the stress-timed and syllable-timed sentences. However, the nature of their distributions is wholly incompatible with the conception of the rhythm typology as encoding a universal, parametric

distinction with strong perceptual manifestations and great importance for acquisition and comprehension. Rhythm is said to contribute to the perception of a foreign accent for non-native speakers, but foreign accents are perceived quickly; there is no need to listen to several dozen sentences to verify that the overall distribution of rhythm is in an inappropriate range. Rhythm has been hypothesized to be crucial for language acquisition by framing the fundamental perceptual unit of speech in a language, as either the stress-foot or the syllable. But how fundamental can such a unit be if it is allowed to switch from one category to the other thirty or forty percent of the time? If ΔC, for example, is in fact a strong metric for rhythm, we are required to believe that a German says 60% of her sentences with a rhythm that would be a natural one in Italian, and that almost any Spanish rhythm could just as naturally be spoken in German.

Few would deny that there is a strong impressionistic difference between the rhythms of languages such as English and Spanish. In light of personal intuitions, and the long and frequently quite passionate and eloquent history of work on the rhythm typology, it is easy to be sympathetic to the argumentum ad populum view that "[t]his typology... must capture some fundamental fact about rhythmic patterns across languages. Otherwise, it would have been relegated to the dustbin long ago"(Nazzi et al., 1998). The impression of rhythmic difference between languages is so strong that when the original basis of rhythm in isochrony proved empirically false, new foundations for the theory were sought, in metrics based on durational variability. But in the quest to find quantifications that replicate the typology rhythm is purported to produce, the definition of rhythm has been narrowed to the point where it is no longer capable of producing the effects that inspired such enthusiasm in the first place.

9 Appendices

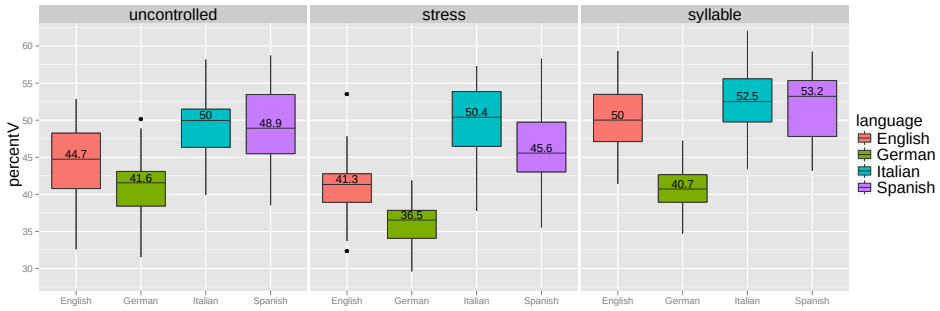

Fig. 3. Boxplot showing *%V* metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

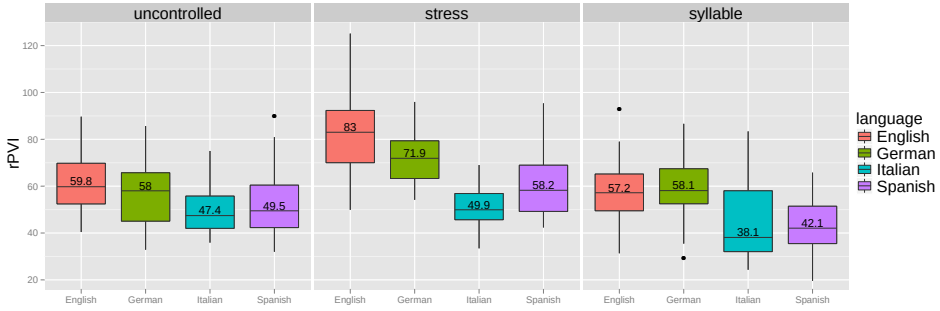

Fig. 4. Boxplot showing rPVI metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

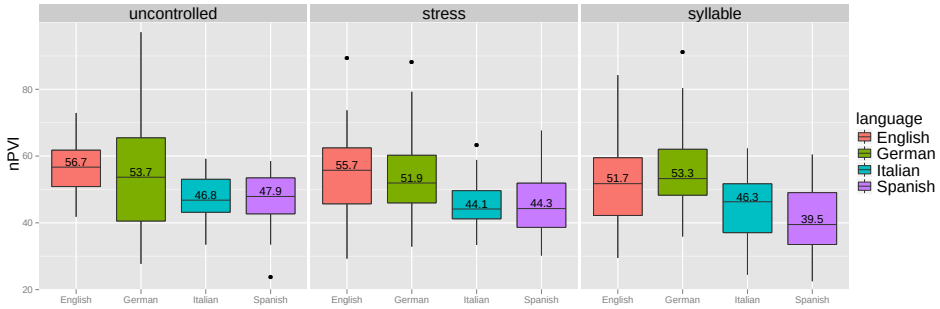

Fig. 5. Boxplot showing nPVI metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

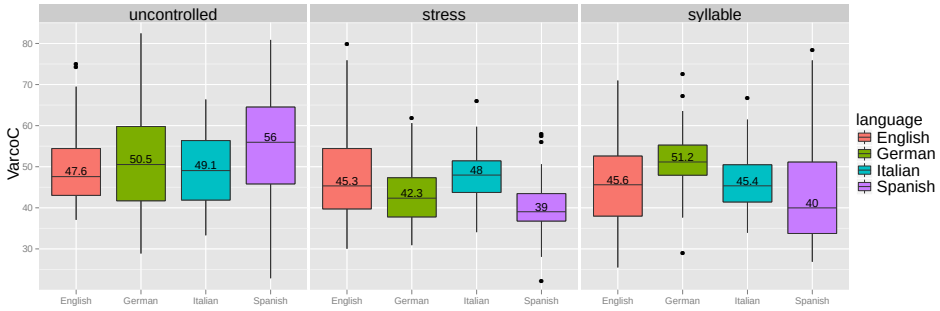

Fig. 6. Boxplot showing VarcoC metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

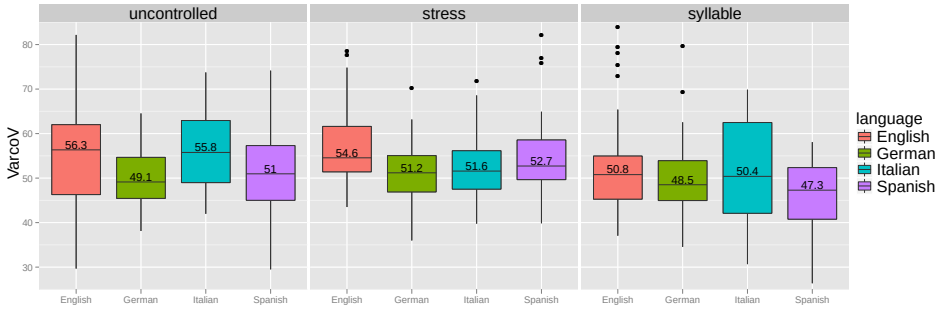

Fig. 7. Boxplot showing VarcoV metrics on uncontrolled, stress- and syllable-timed sentences, for the four rhythmically classified languages

References

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

Arvaniti, A. (2009, January). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*, 47–63.

Arvaniti, A. (in press-a). Rhythm classes and speech perception. In *Prosodies: Context, function and communication.*

Arvaniti, A. (in press-b). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*.

Barry, W., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003, January). Do rhythm measures tell us anything about language type. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2693–2696.

Bertinetto, P. M. (1989). Reflections on the dichotomy ⟨⟨stress⟩⟩ vs ⟨⟨syllable-timing⟩⟩. *Revue de Phonétique Appliquée*, *90 - 91 - 92*, 99–129.

Bertrán, A. P. (1999). Prosodic Typology: on the Dychotomy between Stress. Timed and Syllable-Timed Languages. *Language Design: Journal of Theoretical and Experimental Linguistics*, *2*, 101–130.

Curk, T., Demšar, J., Xu, Q., Leban, G., Petrovič, U., Bratko, I., ... Zupan, B. (2005). Microarray data mining with visual programming. *Bioinformatics*, *21*, 396–398.

Dauer, R. M. (1983, January). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

Dellwo, V. (2006, January). Rhythm and Speech rate: A variation coefficient for âĹĘ C. *Language and language-processing*, 231–241.

Dellwo, V., & Wagner, P. (2003, January). Relations between language rhythm and speech rate. *Proceedings of the International Congress of Phonetics Science*, 471–474.

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, *15*, 87–127.

Grabe, E., & Low, E. L. (2002, January). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*.

Karypis, G. (2003). *CLUTO - A Clustering Toolkit* (Tech. Rep. No. {#)02-017.

Kohler, K. J. (2009). Rhythm in Speech and Language. *Phonetica*, *66*(1-2), 29–45.

Lee, C. S., & Todd, N. P. M. (2004, January). Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, *93*, 225–254.

Ling, L. E., Grabe, E., & Nolan, F. (2000, January). Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, 377–401.

Lloyd James, A. (1940). *Speech signals in telephony*. London: Sir Isaac. Pitman & Sons, Ltd.

Loukina, A., Kochanski, G., Shih, C., Keane, E., & Watson, I. (2009). Rhythm measures with language-independent segmentation. *Proceedings of Interspeech 2009, Sep 6 - 10, Brighton*, 1531–1534.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998, January). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, *24*(3), 756–766.

Nazzi, T., & Ramus, F. (2003, January). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*, 233–243.

Pike, K. L. (1945). *The Intonation of American English, by Kenneth L. Pike*. University of Michigan Publications.

Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. *Proceedings of the 15th international congress of phonetic sciences*, *3*, 337–342.

Ramus, F., Nespor, M., & Mehler, J. (1999, February). Correlates of linguistic rhythm in the speech signal. *Cognition*, 265–291.

Renwick, M. E. L. (2012, February). Quantifying rhythm: Interspeaker variation in %V. *The Journal of the Acoustical Society of America*, *30*, 2567.

Rodriguez, T., & Arvaniti, A. (2012, in prep.). The role of tempo and fundamental frequency patterns in language discrimination.

Rouas, J.-L., Farinas, J., Pellegrino, F., & André-Obrecht, R. (2005, January). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, *47*, 436–456.

Russo, M., & Barry, W. J. (2008, January). Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo. *Proceedings of Speech Prosody*, 419–422.

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

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition).

Russell Horton
russ@bagaduce.com

Amalia Arvaniti
Dept. of English Language and Linguistics
University of Kent