Rhythm measures and dimensions of durational variation in speech[a)]

Anastassia Loukina,[b)] Greg Kochanski, Burton Rosner, and Elinor Keane

Oxford University Phonetics Laboratory,

OX1 2JF,

Oxford,

United Kingdom

Chilin Shih

Departments of Linguistics and East Asian Languages and Culture,

University of Illinois,

Urbana,

Illinois 61801

(Dated: August 15, 2010)

[a)] Preliminary results obtained on part of the corpus were presented in "Rhythm measures with language-independent segmentation", Proceedings of Interspeech, Brighton, 2009, 1531-1534

Abstract

Patterns of durational variation were examined by applying fifteen previously published rhythm measures to a large corpus of speech from five languages. In order to achieve consistent segmentation across all languages, an automatic speech recognition system was developed to divide the waveforms into consonantal and vocalic regions. The resulting duration measurements rest strictly on acoustic criteria. Machine classification showed that rhythm measures could separate languages at rates above chance. Between-language variability in rhythm measures, however, was large and comparable to within-language differences. Therefore, different languages could not be identified reliably. In experiments separating pairs of languages, a rhythm measure that was relatively successful at separating one pair often performed very poorly on another pair: there was no broadly successful rhythm measure. A combination of three rhythm measures was necessary for separation of all five languages at once. Many triplets were comparably effective, but the confusion patterns between languages varied with the choice of rhythm measures. These findings are similar to the results of perceptual studies, and they challenge the reality of rhythm classes.

PACS numbers: 43.70.Kv, 43.70.Fq, 43.70.Jt, 43.72.Ar

I. INTRODUCTION

There is a widespread intuition that languages differ in 'rhythmic structure'. Perceptual studies using processed speech signals support this idea. Even without access to segmental information, both adults and infants can tell apart different languages (for references see Ramus *et al.*, 1999; Nazzi and Ramus, 2003; Komatsu, 2007). Numerous quantitative indices have been offered in attempts to capture the properties of languages that underpin both the intuition and the experimental findings. We follow Barry and Russo (2003) in calling these indices 'rhythm measures' (RMs). Rhythm measures have been widely used for comparison between different languages and varieties (see, for example, White and Mattys, 2007a, for an overview), but they have been subject to criticism for lack of reliability and for dependence on the material (cf Arvaniti, 2009).

The perception of rhythm, however, cannot rest entirely on duration. Barry *et al.* (2009), for example, showed that changes in F_0 as well as in duration influence the perceived strength of rhythmicity. In another study, rate of spectral change proved more important than duration in distinguishing spoken poetry from prose (Kochanski *et al.*, 2010). Since all the current quantitative indices measure duration alone, the term 'rhythm measure' is actually somewhat misleading. Despite its limitations, we retain that well entrenched, traditional term.

In any event, behavioral studies show that people can distinguish between languages when presented with resynthesized signals that primarily contain durational cues. Accordingly, much work has tried to determine which particular rhythm measures are needed to separate languages and varieties (see White and Mattys, 2007a, for an overview). All previous studies have used relatively small corpora of speech. Growing evidence, however, shows that RMs can vary greatly between speakers or texts (Keane, 2006; Arvaniti, 2009; Wiget *et al.*, 2010). A large speech corpus thereby becomes essential for an extensive rhythm study. The corpus should cover numerous speakers and many texts. Furthermore, very few studies

[b] anastassia.loukina@phon.ox.ac.uk

have compared languages on more than two measures, although limited evidence intimates that this may be insufficient for covering cross-language distinctions in rhythm (Ramus *et al.*, 1999; Liss *et al.*, 2009). Indeed, the many available measures should be examined systematically.

A. Purpose of experiment

To meet these requirements, we studied patterns of durational variation by applying fifteen previously published rhythm measures to speech from five languages. The corpus for each language is substantially larger than anything used in past rhythm studies. Three main issues are addressed. First, how well can machine classification identify the languages, using various combinations of the rhythm measures (RMs)? Second, how many RMs are needed to disentangle cross-linguistic variation in rhythm? Third, does the array of the most useful measures depend on the languages being identified?

We use several automatic segmentation algorithms that split speech into consonant-like, vowel-like and silent regions. The algorithms offer uniform, language-independent treatment of acoustic signals, avoiding the inevitable inconsistencies introduced by human labelling and permitting computation of rhythm measures over our large corpus. Finally, machine classification is used to identify different languages from the resulting rhythm measures.

B. Differences between rhythm measures

Several studies have attempted to identify the most useful rhythm measures for separating languages. This effort was partly motivated by evidence that RMs may show great variability between speakers or texts (Keane, 2006; Arvaniti, 2009; Wiget *et al.*, 2010). White and Mattys (2007a) and White and Mattys (2007b) examined which measures best differentiated speech from two languages or from native and non-native speakers of a single language.

Published rhythm measures differ in three respects. Firstly, they use differently determined intervals. Initially, rhythm metrics rested on the durations of vocalic or of consonantal intervals. This reflected the suggestion by Ramus *et al.* (1999) that infants perceive speech as a succession of vowels alternating with periods of unanalyzed noise[1]. In addition, explanations of perceived differences in rhythm have also invoked phonological properties such as vowel reduction and syllable complexity (Dauer, 1983); these properties presumably would affect RMs that employ vocalic or consonantal durations.

More recently, Barry *et al.* (2003) argued that treating consonants and vowels separately forces RMs to miss the combined effect of vocalic and consonantal structure. They therefore suggested that RMs should be defined in terms of syllables or pseudo-syllables. Deterding (2001) had earlier suggested a similar approach, while Liss *et al.* (2009) measured variation in the duration of VC sequences, arguing that these better represent the perception of syllable weight. Recently, Nolan and Asu (2009) suggested further modifications of RMs based on feet.

The second difference between rhythm measures is whether they assume 'global' or 'local' forms. Global RMs capture variation in the duration of particular intervals over an entire utterance. Local measures focus on differences between two immediately consecutive intervals and then average those differences over the utterance. Local measures supposedly differentiate better between various patterns of successive long and short intervals, thereby capturing auditory impressions of rhythm (see Barry *et al.*, 2003).

Thirdly, some measures include a normalization step, while others keep raw durations. Measures based on raw durations are more vulnerable to changes in speech tempo. Normalized measures, however, may level out cross-linguistic differences. Many studies have used a combination of both. For example, Grabe and Low (2002) normalized their vocalic index but used raw values for the consonantal index. In another study, Wiget *et al.* (2010) found that normalized measures of variability of vocalic intervals discriminated best between languages and were most stable under changes in articulation rate. They recommended using a combination of at least two measures that were robust to segmentation procedures and to speech rate.

C. How many measures do we need?

A key question is how many measures are necessary to capture differences in rhythm between languages and dialects. In early work, languages or dialects were mapped onto a two-dimensional plane where the axes represented variation in vocalic and in consonantal intervals, respectively. Later studies used other combinations of measures, selected because of their reliability and/or ability to separate languages or dialects (cf. Barry and Russo, 2003). Most studies used one or two rhythm measures, but there is limited evidence that this may be insufficient to cover rhythmic differences between multiple languages. Ramus *et al.* (1999) found that while just two measures distinguished groups of languages in their corpus, a third pulled Polish apart from English and Dutch. They suggested, however, that this third measure actually reflects phonological properties of the language and may be irrelevant to the perception of rhythm. Recently, a discriminant analysis by Liss *et al.* (2009) indicated that several measures were necessary to distinguish dysarthric from normal speech. These various findings raise the questions of whether just two measures can truly encapsulate cross-language differences in rhythm and whether some particular, limited set of measures can suffice to capture the differences in durational patterns between any two languages or varieties.

D. Segmentation

Before any rhythm measures can be calculated, a first crucial step is the segmentation of the speech signals into vocalic and intervocalic intervals. Unless rhythm measures are defined in a language-independent manner, they cannot be used to compare languages. Accordingly, the definition of ‘vocalic’ and ‘intervocalic’ intervals must avoid any phonological interpretations and ignore syllable and foot boundaries unless they can be approximated in a language-independent way (cf. Ramus *et al.*, 1999).

Most studies on rhythm measures have employed manual segmentation. Its outcome, however, varies between the labellers and, more importantly, depends on their phonological knowledge. Wiget *et al.* (2010) showed that shared criteria amongst labellers produced greater consistency, compared to the variability introduced by different speakers and texts. The authors suggested that human labellers should use an agreed protocol and discuss difficult cases. Even then, manual segmentation has some shortcomings. A labeller's ideas of 'acoustic criteria' often rest on the experience of segmenting English data. When applied to other languages, such ideas may produce counter-intuitive results, prompting re-evaluation (see, e.g., Barry and Russo, 2003; Grabe and Low, 2002). Last, but not least, variability in rhythm measures absolutely requires use of large corpora. Manual segmentation here would be expensive and time-consuming.

Given the problems with manual segmentation, interest has grown in automatic segmentation based purely on acoustics. Galves *et al.* (2002) demonstrated the potential benefits of automatic segmentation and suggested a new set of rhythm measures reflecting the distinction between sonority and obstruency. Similarly, Dellwo *et al.* (2007) argued that languages from different rhythm classes can be distinguished by an analysis of voiced and voiceless intervals. Both studies used small corpora and only global measures of rhythmic variation. Finally, Wiget *et al.* (2010) compared automatic and human segmentation of a small corpus. They first employed automatic forced alignment to match a transcription with the signal. Then they converted the transcription into a sequence of vocalic and consonantal intervals. Scores for traditional rhythm measures computed from the automatic segmentation were within or just outside the ranges produced by the human labellers.

Unfortunately, forced alignment depends upon language-specific transcription and therefore cannot give truly language-independent results. Similar acoustic signals might often be assigned different labels in different languages. Forced alignment, moreover, is insensitive to variation in the realization of individual sounds. For example, it misses lenition and reduction or deletion of segments or syllabic consonants, unless they are reflected in the transcription. In theory, the transcription could be manually adjusted to reflect the connected speech processes, or a comprehensive dictionary of alternative realizations could be compiled. Either alternative would be expensive and would effectively negate the benefits of automatic segmentation.

In this paper, current methods of speech recognition are used to create cross-linguistic statistical models of vocalic and intervocalic regions. Then the models are embedded in segmentation algorithms. To investigate whether different models affect the findings, we compare the results from two segmentation algorithms. We also compare the algorithms against one employing simple acoustic criteria. Finally, all three algorithms and human labellers are compared on the analysis of a sub-corpus of our speech data.

II. METHOD

A. Speech data

Our corpus contained 2300 texts recorded from 50 speakers distributed across Southern British English (N=12), Standard Greek (N=9), Standard Russian (N=10), Standard French (N=9) and Taiwanese Mandarin (N=10). Each speaker read the same set of 42 texts (original or translated) in their own language. These included extracts from "Harry Potter", fables, and the fairy tale Cinderella. On average, texts on contained 217 syllables. Each speaker also read 4 nursery rhymes (75 syllables on average), matched for meter across the languages.

Speakers were 20-28 years old, born to monolingual parents, and had grown up in their respective countries. At the time of the recording, all speakers were living in Oxford, UK. Non-English participants had lived outside their home country for less than 4 years. Recordings were made through a condenser microphone in a soundproof room in the Oxford University Phonetics Laboratory and saved direct to disc at a 16 kHz sampling rate. Texts were presented on a VDU screen in standard orthography for each language. Speakers could repeat any text if dissatisfied with their reading. Overall, 15% of the recordings were repeated, although the fraction varied greatly between speakers. The recordings of each speaker took place in two or three sessions on separate days.

B. Segmentation

We used several different segmentation algorithms (SAs) to split the data into vowel-like and consonant-like components. One algorithm (SA1) depended on loudness and regularity. Two other algorithms used the full spectrum. All SAs acted as recognizers. After the training stage, an SA had no access to transcriptions. It assigned labels according to acoustic properties of the signal.

1. Segmentation algorithm SA1 based on loudness and aperiodicity

For algorithm SA1, we computed time series of specific loudness and aperiodicity (Kochanski *et al.*, 2005; Kochanski and Orphanidou, 2008). These values were smoothed and then compared against thresholds to generate transitions from one segment to another (see fig. 1). The process operated with three types of segments: (1) vowel-like segments with a nearly periodic waveform; (2) segments with an aperiodic waveform which can include frication and/or regions with rapid changes in the waveform; and (3) silences. These three categories are broadly consistent with the specifications of most published rhythm measures, which are defined in terms of vocalic and intervocalic intervals. Five parameters controlled SA1: [a] a smoothing time constant for the loudness and aperiodicity time series (the smoothing process tends to suppress very short segments); [b] the normalised loudness of the silence-to-non-silence transition; [c] the normalised loudness of the opposite transition; [d and e] aperiodicity thresholds for the (2)→(1) and (1)→(2) transitions respectively. Two different thresholds were necessary to prevent small fluctuations in the data from leading to a sequence of very short segments.

An optimization procedure set the parameters for SA1, using a sample of 7143 utterances from the corpus. The sample contained data from 12 speakers, two per language but four for English. It included texts from our corpus as well as short sentences recorded by the same speakers for a larger study. The parameters were adjusted to minimize the mean-squared difference between the number of segments generated by SA1 and the number predicted from

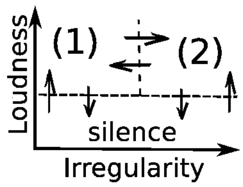

FIG. 1. Transitions between the three states

phoneme-level transcriptions of the utterances. The number of occurrences of segment (1) produced by the SA1 was matched to the number of appearances of vowels and sonorants. The number of occurrences of segment (2) was matched to that of the remaining phonemes. Silences were weakly constrained to appear about 10% as often as the other regions. After this optimization, the parameters were applied across the entire corpus.

2. HTK segmentation algorithms SA2a and SA2b

We developed two other segmentation algorithms, SA2a and SA2b, through the standard HTK toolkit (Young *et al.*, 2006). For SA2a speech was represented as a 26-dimensional standard mel frequency cepstrum coefficient (MFCC) vector. For SA2b speech was represented as 41-dimensional Acoustic Description Vector. The two algorithms also varied in numbers of states, minimal pause lengths, and ways of measuring phoneme duration (see Appendix A for further details). The algorithms were trained on human-segmented data, and the derived models were applied to the whole corpus. Both algorithms are language-independent: no adjustments are made from language to language.

C. Rhythm measures

Using each of the three segmentation algorithms, we computed the 15 rhythm measures (RMs) listed in Table I on each of the 2300 spoken texts in our corpus. Although we use the conventional labels V and C in our designations of these measures, these labels really refer to those assigned by the segmentation algorithms. Our labels reflect the acoustic properties of spoken speech and may not correspond perfectly to phonologically transcribed vowels and consonants. For rhythm measures based on phonological syllables, we used sequences of consecutive consonantal and vocalic intervals. We did not use measures based on VC sequences, which Liss *et al.* (2009) suggested for dealing with syllable weight, since most languages in our corpus do not contrast light and heavy syllables. We also excluded measures based on feet, since the definition of foot is language specific.

Previous studies of RMs have treated pauses and pre-pausal syllables in different ways. To evaluate the importance of such different treatments, we computed each RM in three different ways. First, we calculated the scores for each inter-pause stretch (IPS) then averaged over all IPSs within a text. The average was weighted by the duration of each IPS. Second, we made the same calculation after omitting the final 'syllable'. For each IPS that ended in a vocalic interval, we omitted the final sequence of consonantal and vocalic intervals (CV). For each IPS that ended in a consonantal interval, we omitted the two final consonantal intervals and the intervening final vocalic interval (CVC). Third and finally, scores were simply computed across the whole text, including intervals spanning a pause.

D. Classifier

To quantify variation in RMs between languages, we applied classifier techniques (cf. Kochanski and Orphanidou, 2008). A classifier is an algorithm that decides which language was most likely to have produced an utterance, given one or more observed RMs. We used a Bayesian forest approach (see Appendix B). Insofar as the RMs capture the rhythmic differences between the languages, success or failure corresponds roughly to whether or not a listener could identify the language from its rhythm after hearing a single spoken paragraph.

Our linear discriminant classifiers assume that the log likelihood ratio between the probabilities of any two languages is a linear function of the input rhythm measures. The classification boundaries for each language then form a convex polygonal region in the space of the observed RMs.

Classifiers were built with 12 different combinations of non-overlapping training and test sets. These sets came from a typical 3-to-1 split of the dataset, respectively. The algorithm was executed from 20 different starting points for each training-test combination. We report averages of the resulting 240 (12*20) individual runs on each dataset. The variation in performance from one instance of a classifier to another is used to determine whether differences in performance are statistically significant. To prevent classifiers from learning patterns of individual speakers, data from a given speaker never appeared in both the training and the test set for a given run.

III. RESULTS

For each segmentation algorithm and pause condition, we ran 15 classifiers based on single RMs, 105 based on all possible combinations of two RMs, and 455 classifiers based on all possible combinations of three RMs. Finally, one classifier used all 15 RMs for a grand total of 576 classifiers. For each of the three SAs, then, there were 1728 runs that included different pause conditions. Similarly, each pause condition was represented in 1728 runs that included different segmentation algorithms. We first applied all classifiers to pairwise identification of all 10 pairs of languages in our corpus. We then repeated the analysis for the whole corpus at once, testing how well the classifiers separated all five languages.

Our results are based on the probability of correctly identifying the language of a paragraph. If this is large (i.e. near 1.00), it means that the data from the various languages can be separated into distinct groups by straight lines.[2] One can think of this as a test of the hypothesis that different languages form separable clumps. A small identification probability (e.g. near chance) happens if the data from different languages are intermingled.

One complication is that the chance level varied between runs on only two languages at a time and those using the whole corpus. We defined our chance level conservatively, to be the best possible performance of a classifier that knows the relative frequencies of the classes, but not the RM value(s) for a particular paragraph. The chance level is then the proportion of passages from the most frequent language in the training set. This varies from experiment to experiment, and even for the different classifier instances within a forest, since the various training sets do not have exactly the same composition.

To allow simple comparisons, we report both the proportions of correct identifications, $P(C)$ and a figure of merit designated K. This is computed as $K = \frac{P(C)-chance}{1.00-chance}$, where 1.00 represents perfect performance. Thus, K varies between 0 for classifiers that perform at chance and 1 for perfect classifiers. We used z-tests to assess both the significance of differences between $P(C)$ and chance for each classifier and the significance of differences between classifier performances. Since we foresaw a large number of tests, we set the significance level of the tests (α) conservatively at .01.

A. The effect of segmentation and computation method

The three methods of handling pauses in the computation of rhythm measures did not affect classifier performance. For classifiers using identical RMs, differences in P(C) between pause conditions were significant in less than 1% of all 5184 (3*1728) possible pairwise comparisons. These results are at chance. Segmentation algorithm did not influence the performance of classifiers that treated all five languages in one run. Classifiers using identical RMs yielded significantly different values of P(C) between two SAs in less than 1% of the 5184 possible comparisons. This again is at chance.

In contrast, for classifiers that sorted just a pair of languages, significant differences in P(C) appeared between SAs in about 2% of the 51840 cases. (Each of the 10 possible pairs underwent 5184 comparisons across SAs.) Across all pairs, the differences mainly occurred between SA1 as against the two HTK-based algorithms, SA2a and SA2b. Generally, SA1 performed worse than the other two. The $P(C)$ was higher for SA2a than for SA1 in 70% of significantly different comparisons. Where $P(C)$ differed significantly between SA2b and SA1, the former performed better 93% of the time.

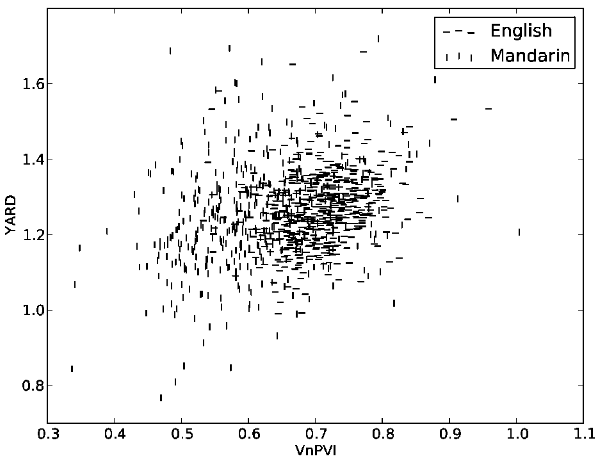

FIG. 2. The values of of VnPVI and YARD in English and Mandarin data (K=.46, P(C)=.76, chance=.54)

In short, computation procedure had no effect on the accuracy of language identification. Segmentation algorithm had only a small effect. Accordingly, the classifier results presented below rest on algorithm SA2a and on the computation of RMs that omits the final sequence of consonantal and vocalic intervals (CV or CVC) in each IPS.

B. Classifiers for pairs of languages

Most behavioral language identification experiments have used pairs of languages. Therefore, we first took the 576 classifiers using all possible combinations of 1, 2, 3 and 15 RMs and applied them to the 10 possible pairs of languages in our corpus. Languages in all pairs could be separated above chance, but P(C) never reached unity (perfect identification). That occurred because all pairs of languages showed substantial overlaps in the values of rhythm measures. Fig. 2 shows the distribution of values for a randomly selected pair of measures and pair of languages.

For 9 of the 10 pairs of languages, maximum P(C) could be achieved using only one RM (see Table II); adding further RMs gave no significant gain in P(C). The one exceptional case was identification of Russian vs Greek, where three RMs were needed to maximize $P(\mathrm{C})$. Table II shows that the most successful single RM (or successful set of RMs) depended on the language pair. For example, a single consonantal measure allowed the best possible separation of Mandarin from English, but no consonantal measure could separate Mandarin from French above the chance level.

Not all pairs of languages showed the same degree of confusion. Mandarin was identified most readily: across all four language pairs that included Mandarin, the best classifiers gave an average K of .53 and average $P(\mathrm{C})$ of .78 (average chance .53). In contrast, across all six pairs of European languages, the best classifiers yielded an average K of .30 with an average $P(\mathrm{C})$ of .68 (average chance .54). The difference in $P(\mathrm{C})$ between any two pairs of European languages was not significant.

C. Identification of all five languages

1. The success of individual RMs

Only eight of the 15 measures performed above chance in correctly sorting all five languages in our corpus. These were the two ratio measures, all normalized vocalic measures and all normalized CV-based measures (see Table I). Their average $P(\mathrm{C})$ was .33 (chance=.23, K=.12). No significant differences appeared amongst these eight classifiers.

2. Classifiers based on several RMs

We reported above that the RM necessary to achieve the maximum observed $P(\mathrm{C})$ for a pair of languages differed between pairs. As these results imply, classifiers required a combination of RMs to achieve maximum $P(\mathrm{C})$ for all five languages. Given a single CV or V measure, adding two more measures from the V, CV, or ratio types raised $P(\mathrm{C})$ to an average .44 (K=.27). For classifiers based on a single ratio measure, significant improvement required adding three V or CV measures. Average $P(\mathrm{C})$ for classifiers using four such RMs then reached .46 (K=.30). Finally, although no classifier using a single consonantal measure had performed above chance on all five languages, a combination of C-based measures improved matters significantly. Fifteen classifiers based on pairs or triads of local and global C-based measures achieved an average $P(\mathrm{C})$ of .36 (K=.17).

Beyond this, adding more RMs to the classifiers led to no further gains. No classifier correctly identified the five languages all the time. Indeed, classifiers based on all 15 rhythm measures gave a *P(C)* around .55 (K=.41) of data. This does not differ significantly from the rates achieved by virtually all classifiers that used just three vocalic or syllabic measures. In short, within the 15 RMs studied here, maximum accuracy in identifying our five languages required a combination of just three of the right types of measures. Moreover, many sets of 3 RMs performed similarly well.

D. Relations between RMs

Many of the published RMs rely on similar calculations and therefore are highly correlated. To estimate the minimum number of RMs needed to cover the variation between our five languages, we performed multidimensional scaling (MDS) with PROXSCAL. To create a dissimilarity measure, RMs were intercorrelated within each language. Then 1-r^2, where r is a correlation, gave an ordinal dissimilarity measure between two RMs, ranging from 0 to 1. Languages were treated as separate sources for PROXSCAL.

The dissimilarities between the 15 RMs gave rise to a 5-dimensional solution (stress=.008). The dimensions seemed to represent distinctions between subgroups of RMs due to type of interval (C, CV, ratio, V) and to presence or absence of normalization. No grouping appeared that reflected scope (local or global). Languages differed modestly in their weights on the different dimensions.

At first sight, this seems to disagree with the fact that more than 3 RMs give no significant improvement in *P(C)* for classification of five languages. The MDS solution, however, addresses language identification only indirectly: it shows that with just 5 RMs one can accurately predict any of the other ten. This sets a maximum to the number of RMs that

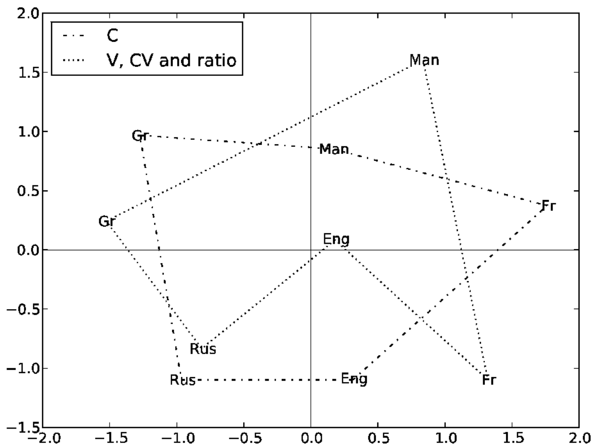

FIG. 3. The location of the 5 languages on MDS map computed using classifiers based on C-only measures or V/CV/ratio measures (V: stress=.18, R^2=.9, C: stress=.18, R^2=.9)

could be used to identify languages. The success of 3-RM classifiers readily fits with this: there are 5 independent RMs, but languages in our corpus only differ in three of them. It should also be noted that the criterion of "statistically significant improvement" used in this study is conservative: a fourth RM might yield some small improvement that could be detected with an extremely large experiment.

E. Relative location of languages

Although classification of all five languages using different combination of RMs can yield comparable values of P(C), the confusion patterns depend on the types of the RMs. To demonstrate this, we created two 5 X 5 asymmetric matrices containing proportions of confusion between pairs of languages. One matrix held the results for all significant 3-RM classifiers based on on CV, V and ratio only measures. The other was constructed from the results of all significant 3-RM classifiers using C measures only. Fig. 3 shows the two MDS maps produced by ALSCAL from these matrices. The two maps reveal different confusion patterns. Classifiers based on V/CV/ratio measures separated Mandarin from the European languages but often confused those four. Classifiers using consonantal measures exclusively put Russian and English in one MDS region and French, Greek and Mandarin in another. This agrees with the finding that consonantal measures separate Mandarin pairwise from Russian or English but not from Greek or French. For both maps, however, the matrices revealed substantial confusion between each pair of languages.

F. Comparison between automatic and manual segmentation

Our automatic segmentation algorithms were designed to avoid the language-specific aspects of human segmentation of speech. Human segmentation inevitably reflects knowledge of the language being labelled. Furthermore, automatic segmentation is inherently consistent and reproducible, while human segmentation is not. Thus automatic and human segmentation could never be expected to agree entirely.

Manual segmentation of our large corpus is impractical. Nevertheless, an important theoretical question remains: Does human segmentation agree well enough with automatic segmentation so that it would yield our basic findings on RMs and on identification of different languages? A segmentation algorithm suitable for quantifying rhythm would agree well with human labels on longer vowels and obstruents, but it might well disagree on less clear-cut cases such as sonorants or short vowels.

In order to get some grasp on this question, we chose a test set of 30 spoken paragraphs from our corpus, covering all five languages. They were segmented by trained phoneticians in the same way as the set originally used to train the SAs. One author segmented 2 paragraphs. The remaining 28 were segmented by seven other phoneticians, each from a different institution.

To compare SAs against the phoneticians, we applied each of our three segmentation algorithms to the test sub-corpus. We divided the human labels into four broad categories: vowels, voiced obstruents, voiceless obstruents and sonorants. At each 10 ms epoch of speech, we recorded both the broad human label and the automatic label of 'V', 'C' or 'S'. Then within each language, we computed the percentage of co-occurrences of 'V', 'C' or 'S' with each broad human label.

We first present detailed results for the comparison with SA2a, and then we consider differences between SA2a and the other two algorithms. Segmentation algorithm SA2a treated human labels for vowels or consonants as 'S' on less than 1 per cent of all occasions. This mainly arose from differences in placement of phrase-final and phrase-initial boundaries and from occasional differences in segmentation of voiceless plosives. Otherwise, agreement on identification of silences was almost perfect. We therefore dropped silences from further analysis.

Table III gives the percentages of 'V' and 'C' labels assigned to each of the four broad human labels. Of the epochs labelled as vowels by the phoneticians, 88%- 93% were tagged automatically as 'V', depending on the language. The bulk of the disagreements concerned the high vowels [i], [u], and [y] and the unstressed [ə]. Less than 85% were tagged by SA2a as 'V'.

Voiceless obstruents were treated by SA2a as 'C' in 83-89% of the samples. More serious disagreement appeared on the English [h], with a 61% tagging as 'V'. (This agrees with the view that in English, and possibly in other languages, [h] is acoustically closer to approximants than to other fricatives (cf. Ladefoged and Maddieson, 1996, p. 326)). As expected, agreement between human and automatic segmentation was worse for sonorants and voiced obstruents than for voiceless obstruents. Sonorants were generally recognized as 'V' (77-91%). Voiced obstruents showed the greatest discrepancies, with the voiced fricatives [v], [ð], [ɣ] often recognized as vowels.

In short, segmentation algorithm SA2a successfully identified most unreduced vowels as 'V' and most voiceless obstruents as 'C'. It had learned the difference between more and less sonorous segments, and it apparently applied criteria similar to those used by phoneticians.

Algorithm SA2b used feature vectors rather than the MFCC vectors implemented in SA2a. The former consistently tagged vowels and voiceless obstruents as 'V' (90-94%) and as 'C' (79-89 %), respectively. Likewise, sonorants were mainly marked as 'V' (78-88%).

Voiced obstruents showed the most variation, with patterns of tagging similar to those for SA2a.

Algorithm SA1 employed simple acoustic criteria. It mapped automatic labels onto the four human categories slightly less consistently than its two more complex partners. Only 79-88% of vowels labelled by humans were marked as 'V', while 79-89% of voiceless obstruents were tagged as 'C'. A noticeable difference appeared in the tagging of voiced stops as 'C' between Russian, Greek and French versus English (20-30% vs 63-79%). This reflects differences in the acoustic correlates of phonological voicing.

Finally, we recoded the labels assigned by the phoneticians into three categories of Vowel, Consonant, and Silence. The sonorants were coded as vowels. We treated labels at each 10 ms epoch as separate observations, giving 4000-7000 observations for each test paragraph. We excluded initial and final silences where they were labelled by both sources. Cohen's kappa was then used to compare the automatic tags of 'V', 'C', and 'S' against the three recoded categories of human labels. This statistic measures overall agreement between automatic and manual segmentations.

The box plots in Fig. 4 display the values of kappa for agreement between each segmentation algorithm and the phoneticians. The median kappa value for both SA2b and SA2a is about 0.75. This is interpreted as 'excellent agreement beyond chance'. The median kappa of 0.65 for SA1 suggests fair to good agreement (see Banerjee *et al.*, 1999, for further discussion of the use of kappa).

IV. DISCUSSION

In order to study the nature and complexity of patterns of durational variation as reflected by rhythm measures, we analyzed a large corpus of speech from five languages. Automatic segmentation and machine classification were necessary to do this. We discuss our three major findings in order.

First, we found that languages differed in durational patterns. Within-language varia-

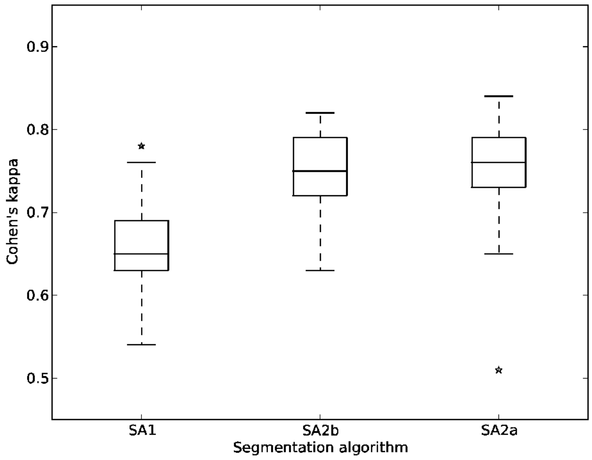

FIG. 4. The values of Cohen's kappa between automatic and manual segmentation. The automatic labels for each paragraphs were compared to all available manual labels. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to the most extreme data point within 1.5 IQR of each of the quartiles. Outliers are indicated by asterisks.

tion is so high, however, that it would appear impossible to identify one language reliably from rhythm measures based on single paragraphs. This conclusion agrees with studies on human language identification. Numerous studies show that when listeners confront a processed signal lacking segmental information, they cannot identify the originating language with perfect accuracy. The exact success rate depends on experimental conditions and on the languages. In studies with low-pass filtering of speech from two languages, $P(\mathrm{C})$ for identification is around .63-.68, above chance of .50 (see Komatsu, 2007, for references). Our classifiers were at least as accurate as that in identifying two languages.

Our second main finding is that no single RM or set of RMs was the best at identifying all pairs of languages. The most effective choice differed from one language-pair to another. In agreement with results reported by White and Mattys (2007a) and Wiget *et al.* (2010), normalized vocalic measures were generally more successful than consonantal or non-normalized measures. However, our results imply that a search for the 'best' RMs for separating all languages is fruitless.

Indeed, it is hardly surprising that different measures are optimal for identifying the members of different pairs of languages. Variation in duration is a product of many factors; among them are stress, syllable complexity, realization of individual sounds, and differences in sentence prosody, in speech tempo and in subject-specific patterns. One or two relatively simple measures would seem very inadequate for capturing all possible differences in rhythm between languages. Even if the measures were fine-tuned to capture one specific contrast, such as temporal stress contrast (cf. Wiget *et al.*, 2010), other factors would still vary across other pairs of languages/varieties.

Our third main finding is that three measures were necessary to achieve optimal identification of all our five languages at once. Given that no single measure led to optimal separation across all pairs of languages, it is only logical that several measures should prove necessary to achieve maximum identification of more than two languages. Once again, the maximum $P(\mathrm{C})$ achieved by the classifiers identifying five languages is comparable to results from human identification. For example, Navrátil (2001) reported a $P(\mathrm{C})$ of .49 (chance=.20) for identifying German, English, French, Japanese and Chinese. The number of measures necessary to separate the members of other corpora might depend on the languages represented.

Different patterns of grouping occurred among our five languages, depending on the choice of rhythm measures. The debate around RMs has often been linked to the concept of rhythm class, to the distinction between stress-timed and syllable-timed languages and to the question of whether languages form discreet classes or a continuum. The historical background of this debate and the arguments on both sides have been widely discussed (see for example Ramus, 2002; Keane, 2006). The 'rhythm class' concept predicts that languages within the same class will overlap on rhythm measures. We observed such overlap. The concept, however, also requires constant grouping patterns: languages from the same rhythm class should always show greater confusion than languages from different classes. We found no such constant pattern of confusion within our five languages. Different combinations of measures produced different patterns of confusion between particular languages or particular subgroups of languages.

This absence of a consistent confusion pattern fits with the fact that the single RM yielding maximum identification for two particular languages shifted, depending on the pair of languages. Furthermore, two languages that showed similar values on some RMs could still be separated on others. The pattern of effective and ineffective RMs varied across pairs. Conclusions about the similarity of two varieties based solely on rhythm measures seem to depend largely on the choice of measures and on the expectations of the researcher (cf. also Arvaniti, 2009, for similar remarks).

Perceptual studies have been offered as evidence for the rhythm class concept. The listener's native language, however, seriously influences the results of such studies, just as it plays a crucial role in speech segmentation (cf. Murty *et al.*, 2007; Tyler and Cutler, 2009, and references therein).[3] Experiments with processed signals reveal that both infants and adults are generally better at distinguishing their native language from a foreign language than at distinguishing between two foreign languages. These studies provide little evidence for consistent grouping into rhythm classes (cf. also Arvaniti and Ross, 2010, for critical review of other studies).

For example, Nazzi *et al.* (2000) reported that 5-month-old American infants discriminated languages traditionally assigned to different rhythm classes such as Italian and Japanese. They also discriminated languages traditionally assigned to the same rhythm class if one language was English but not when both languages were foreign. Ramus *et al.* (2003) found that French students could only discriminate at chance between processed Spanish and Catalan stimuli. In contrast, Bosch and Sebastian-Galles (1997) reported that 4-month-old Spanish and Catalan infants discriminated low-pass filtered versions of speech from the two languages. Similarly, Szakay (2008) found that listeners who were highly integrated into either of two ethnic communities were better at discriminating processed signals representing the two ethnolects than were less integrated listeners.

Listeners apparently use different acoustic cues to discriminate between languages, and the cues depend on the listener's native language or familiarity with the languages being tested. This undermines the use of perceptual results to buttress any particular grouping of languages into classes. The reality of such classes becomes questionable.

Besides our main study with automatic segmentation and identification of languages, we compared automatic and manual segmentation of the same (necessarily limited) set of texts. Excellent but not perfect agreement was found between the labels from the two sources. The results have two important consequences. First, it once again shows that segments placed by a human labeller in the same phonological category may be assigned automatically to different categories based purely on acoustic properties. Hence, manual segmentation cannot be language-independent. Human labellers know the interrelationships between acoustic properties of the speech signal and phonological contrasts. Consequently, rhythm measures based on manual labelling are subject to the influence of language-specific phonological interpretations, making it difficult to achieve language-independent segmentation. Second, our comparison of human and automatic labelling suggests that perception experiments using substitution of segments (for example, substituting [s] for all consonants and [a] for all vowels) reflect the investigator's own prior phonological interpretations. Future experiments should employ signals with gradient transitions between more and less sonorous synthetic segments.

V. CONCLUSIONS

On average, the languages that we studied with a language-independent segmentation procedure proved to have their own particular patterns of durational variation ("rhythm"). However, there is substantial variation within each language on every RM. Because of this variation, one cannot reliably identify a language or determine its properties from published duration measures computed from a single paragraph.

The differences between the five languages in our corpus cannot be captured by only one rhythm measure. While most pairs of languages could be separated fairly well with a classifier based on just one carefully-chosen RM, different pairs needed different RMs. This suggests that languages differ rhythmically in a variety of ways.

Combinations of three RMs were needed to reach the highest correct identification rate for all five languages at once. These findings and multidimensional scaling show that linguistic rhythm is a multidimensional system. However, there are many different combinations of three RMs that are nearly equally effective. Overall, our machine classifier results are as accurate as human identification of languages in perception experiments.

Our results are not consistent with the traditional rhythm class hypothesis that would put our languages into two (or three) sharply-defined classes. The rhythm class hypothesis implies that many combinations of RMs would give the same groupings of languages. Our data show that languages group differently, depending upon which rhythm measures are used to classify them. Plausibly, each rhythm measure captures different language properties.

Finally, human segmentation of a small sub-corpus of speech agreed well with the labels produced by applying our segmentation algorithms to that sub-corpus. There were systematic differences, however, showing that manual labelling of speech depends on phonological interpretations. Therefore, experiments that compare manually-obtained durations across two or more languages have an intrinsic confound: they simply cannot distinguish differences between languages from language-dependent differences in the segmentation process.

Acknowledgments

This project is supported by the Economic and Social Research Council (UK) via RES-062-23-1323. The authors would like to thank John Coleman for useful discussions. We acknowledge the National Science Foundation for providing support to Dr. Shih via IIS-0623805 and IIS-0534133. We also thank Speech Technology Center Ltd. (St.-Petersburg, Russia) and Institute for Speech and Language Processing (Athens, Greece) for their help with automatic transcription of the data.

APPENDIX A: HTK-BASED SEGMENTATIONS

Segmentation algorithms SA2a and SA2b were developed using the standard HTK toolkit. Segmentation algorithm SA2a uses three labels, Consonant, Vowel, and Silence, that correspond to spoken consonants, vowels, and silences, respectively. The Silence label captures silences at the end of each utterance and between phrases. As a final step in the processing, the algorithm merges runs of consonants and of vowels into consonantal and vocalic regions, respectively.

The acoustic model for Consonant contains four alternative, mutually independent sub-models, each roughly representing a major group of spoken phones. Each sub-model is a 3-state sequence, with looping allowed. A state corresponds to a relatively steady part of the phone: for example, it might detect the moment of closure of a variety of stop consonants. All consonant states share the same diagonal variance. Consonants are a minimum of 30 milliseconds long. The Consonant model was trained on individual consonants, so when it met a consonant cluster, it often recognized several consonants in sequence.

The Vowel model uses six 3-state sub-models. It also has another 36 sub-models designed to identify diphthongs. A diphthong sub-model consists of the initial and middle states of one vowel sub-model, then the middle and final states of another. It therefore is four states long and shares states with the vowel sub-models. Finally, the Silence model is at least 100 milliseconds long. This prevents it from responding to short closures that may occur in stop consonants. One hundred milliseconds corresponds roughly to the boundary between short silences that often go unnoticed by listeners and longer ones that are explicitly interpreted as pauses. The Silence model is constructed from two 3-state and two 4-state sub-models that can follow each other in any combination, so trajectories pass through multiples of ten states.

Algorithm SA2a was trained on 19 human-segmented spoken paragraphs. Four professional phoneticians, including three of the authors, independently labelled data in their native language or in a language in which they were reasonably fluent. They used broad phonetic transcriptions and were only given standard guidelines. The labels assigned by phoneticians were recoded into three categories of Vowels, Consonants and Silences. The sonorants were recoded into vowels. The final training data contained 9793 segments (61% English, down to 2% Mandarin) that included sufficient admixtures of each language to allow construction of Gaussian mixtures of English and non-English.

The SA2a algorithm was trained once to establish a rough system. Then the middle state of each Consonant submodel was broadened to include a second mixture, and SA2a was retrained. This brought the complexity of the consonant models closer to that of the vowel models. The retrained SA2a was finally used as a recognizer on entire corpus of speech data exclusive of the training data. Speech was represented as standard mel frequency cepstrum coefficient (MFCC) feature vectors (13 components + derivatives). A grammar put two constraints on recognition: first, the sequence of phones that represent an utterance must start and end in a silence; and second, two immediately successive silences are prohibited.

Segmentation algorithm SA2b generally follows SA2a but with several changes. Sequences marked by the phoneticians as entirely of consonants were mapped into a single segment before training. Likewise pure sequences of vowels in the training utterances were mapped first into a single vocalic segment. For Consonant and for Vowel, SA2b has only two sub-models each, and each of these has three states. The Silence model has a minimum length of 130 milliseconds. It consists of a single sub-model that allows backward steps of 20 ms to 80 ms. It thereby can avoid confusion by substantial, complex repetitive structures within silences such as breathing noises and lip smacks.

Like SA2a, the SA2b algorithm was trained once on the human-segmented spoken paragraphs to establish a rough model. Then the middle state of each Consonant and Vowel submodel was modified to include four mixtures. Four selected states in the silence model were also enhanced to four mixtures. After these alterations, SA2b was re-trained and finally used as a segment recognizer on the corpus of speech. Audio processing for SA2b employed a 41-dimensional Acoustic Description Vector as against the 26-dimensional MFCC+derivatives used in SA2a. This larger vector gives somewhat more emphasis to spectral shape and uses only 5 components of derivative information.

The grammar for SA2b requires an alternation between C and V segments, with occasional silences. So the algorithm must try to model a complex consonant cluster with a single phone. In contrast, SA2a can use several Consonants in sequence to represent a consonant cluster. This is a substantial difference. It forces SA2b to represent a potentially very complex consonant cluster with a single model that was limited to three states. Unlike algorithm SA2a, SA2b needed no final stage to merge repeated pairs of consonants or repeated vowel pairs. It also was subjected to the same two constraints on treatment of silences as was S2a.

APPENDIX B: BAYESIAN FORESTS OF LINEAR DISCRIMINANT CLASSIFIERS

Each "classifier" used in this paper is actually a group of 240 closely related instances. This is a classifier forest approach, inspired by Ho (1998). When applied to small data sets, a forest has the advantage of reporting partial success as well as reporting an item as correctly or wrongly classified. Partial success occurs when some classifiers in the forest identify the item correctly while others treat it incorrectly; this reduces statistical noise compared to using a single classifier.

More importantly, a forest provides a better assessment of how accurately the classifier boundaries are known. Conventional classifiers often report class boundaries, half-way between the outliers of each class, as if they were precisely known. A Bayesian forest samples all plausibly good classifiers. Hence, the variation in boundary positions reflects the true uncertainty about the underlying boundaries. Finally, the various classifier instances can be combined into a ensemble classifier that potentially generalizes to new data more reliably than a single classifier (cf. Tumer and Ghosh, 1996).

The classifier forest is generated in two steps. First, the data are randomly split into a training and a test set. Successive splits are anti-correlated, making the number of times each item is chosen for a test set more uniform than expected from independent random splitting.

Second, for each test-set/training-set split, a bootstrap Markov Chain Monte Carlo (BMCMC) optimizer and sampler (Kochanski and Rosner, 2010) generates linear discriminant classifiers that individually separate the data into N classes as well as possible. Each classifier is a sample from the distribution of all classifiers that are consistent with the training set. (The BMCMC routine is implemented in the `stepper` class in `mcmc_helper.py` and `BootStepper` in `mcmc.py`; these are available to download at Kochanski (2010b).)

In a linear discriminant classifier, each class i has an associated likelihood function:

$$L_i(\vec{c}) = \vec{c_i} \cdot \vec{v} + \alpha_i, \tag{B1}$$

where $\vec{v}$ is the position at which evaluation is occurring, $\vec{c_i}$ are coefficients that describe the class, and α_i relates to the overall preference for class i. (Class i is a particular language in our case.) The probability of assigning a given datum to class i is

$$P_i(\vec{c_i}) = L_i(\vec{c_i}) / \sum_i L_i(\vec{c_i}). \tag{B2}$$

(The final α_i and $\vec{c_i}$ can be both set to zero without loss of generality, which we do.)

The probability density of sampling a particular $\vec{C}$, where C represents complete classifier forest, is the Bayesian posterior probability, given the training data:

$$P(\vec{C}) \propto \prod_j P_d(\vec{c_j}). \tag{B3}$$

Here, j runs over all the training data, and d is the index of the correct class for each datum. In this algorithm, we use a prior that assigns equal probability to each class, and all the measurements are assumed to be mutually independent.

This is a model that does not have sharp class boundaries. Rather, at each point, there are probabilities that the datum could be a member of any of the classes, and these probabilities change smoothly. (Though the model can represent cases with sharp class boundaries by making the change very rapid.)

The BMCMC sampler uses a bootstrap version of the Metropolis algorithm (Metropolis *et al.*, 1953). The algorithm keeps track of the current value of $\vec{C}$ and attempts to change it at each step. A change that increases $P(\vec{C})$ is accepted, and $\vec{C}$ is moved to the new position. A change that decreases $P(\vec{C})$ is accepted with probability $P(\vec{C}_{\text{new}})/P(\vec{C}_{\text{old}})$. Equation B3 is written as a proportionality, because the denominator of Bayes' Theorem is an impractical multidimensional integral that (fortunately) is independent of $\vec{C}$; this independence allows computation of the step acceptance probability without the need to integrate.

The BMCMC algorithm generates changes by making steps proportional to differences amongst an archive of its previous positions. It is described more fully in Kochanski and Rosner (2010); it has been used in prior work, notably Alvey *et al.* (2008) and Braun *et al.* (2006), and is available for download (Kochanski, 2010b). It is first run to convergence (via `stepper.run_to_bottom` in `mcmc_helper.py`) and then run to generate (in this instance) 20 samples of $\vec{C}$ from the distribution of classifiers for each test/training split. (via `stepper.run_to_ergodic` in `mcmc_helper.py`). These samples are chosen with a probability that reflects how well Equation B3 matches the available data; thus most samples will come from the vicinity of the maximum likelihood classifier.

One can define confidence regions from these samples. In particular, if the actual data are generated from Equation B2, there is a 95% chance that the underlying parameters used to generate the data will lie within a confidence region that contains 95% of the generated samples.

The classifications that the algorithm produces (and the class boundaries) are simply the class that gives the largest probability in Equation B2, or (equivalently) the maximum likelihood class (Equation B1). Class boundaries are convenient for visual display. More importantly,a "hard" classification is useful because it leads to a good (and easily understandable) measure of the classifier performance: the probability of correct classification.

The work here used the `qd_classifier` program with the `-L` flag to produce linear discriminant classifier forests. The `-group` flag was used to extract the speaker ID, making the classifier group data by speakers. The classifier code is available for download (Kochanski, 2010a). Related code, `l_classifier`, is also available and recommended for items that are nearly independent.

In `qd_classifier`, the data are split into test and training sets via the `bluedata_groups` class in `data_splitter.py`. (We use 12 splits in this work.) This splitting is a two-pass algorithm and is a stratified sampling scheme. First, we assign a group (a subject in our case) to either the test or the training set. This assignment is anti-correlated with previous assignments. For example, if in previous splits, subject D3 has not yet been assigned to the test set, D3 is more likely to be assigned this time. Then, the algorithm samples (without replacement) from each speaker, so that the test and training sets have nearly the same fraction of items from each class. This sampling is also done in an anti-correlated fashion, so that all items will be in the test set nearly the same number of times.

This procedure insures that data from a given speaker never appear in both the training and test set. A classifier's success rate therefore does not measure its ability to learn the quirks of any individual speaker. Rather, it measures only the properties shared by the entire sample of speakers.

Our hybrid scheme of using multiple training/test-set splits combined with a Bayesian sampling of classifiers with each training set is well-behaved even in cases where there are only a few groups. For instance, in a data set with only four groups (e.g. four experimental subjects), there are only four ways to make a split into a training set and a test set that hold 75% and 25% of the data, respectively. If more than four samples are needed, e.g. to compute error bars for the probability of correct classification, the Bayesian procedure can still generate multiple samples from each training set.

Multiple test-set/training-set splits are valuable, because real data are probably not generated from Equation B2 and utterances are generally not independent. Properties of utterances can be correlated with each other for many reasons, but the most common and often the most important one is that the same person generates them. If each individual has a different voice or style of speech, inter-speaker variation can be much larger than the variation within an individual's utterances. In such a case (as here), two utterances from the same speaker are not independent because one can use the properties of the first to predict the properties of the second.

If utterances are not independent, samples drawn from a BMCMC sampler based on Equation B3 will give an overly narrow distribution of $\vec{C}$, because Equation B3 falsely assumes independence. In an extreme case where inter-speaker variation dominates and there are many utterances per speaker ($N_{\text{ups}} \gg 1$), error bars would be underestimated by a factor of ${N_{\text{ups}}}^{1/2}$, causing false significances in hypothesis tests.

This problem is germane to all work where statistical tests do not account for inter-speaker variation; many published papers suffer from it, not just Markov chain Monte Carlo samplers. Our solution is to compute a new group of BMCMC samples for each test/training-set split. Each split is approximately a bootstrap (Efron, 1982) sample of speakers, thereby capturing the inter-speaker variation. Within each split, the BMCMC sampler reflects intra-speaker variation, and the overall result reflects the full variability of speech.

ENDNOTES

1. Gerhardt *et al.* (1990) measured the intrauterine acoustic environment of fetal sheep. They found that high frequencies are somewhat attenuated, but with only a single-pole filter. As a result enough high frequency information remains so a fetus could potentially discriminate among the consonants or among the vowels.

2. More generally, by N-1 dimensional hyperplanes for a N-dimensional classifier.

3. The effect of native language on the perception of rhythm even extends beyond the domain of speech. In their study of the perception of rhythmic grouping of nonlinguistic stimuli by English and Japanese listeners, Iversen *et al.* (2008) showed that language experience can shape the the results.

REFERENCES

Alvey, C., Orphanidou, C., Coleman, J., McIntyre, A., Golding, S., and Kochanski, G. (**2008**). "Image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: a comparison using automated image processing", Int. J. CARS **3**, 457–464.

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

Arvaniti, A. and Ross, T. (**2010**). "Rhythm classes and speech perception", in *Proceedings of Speech Prosody 2010, Chicago*, 100887: 1–4.

Asu, E. and Nolan, F. (**2005**). "Estonian rhythm and the pairwise variability index", in *Proceedings of FONETIK 2005, Göteborg, 25-27 May 2005*, 29–32.

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (**1999**). "Beyond kappa: A review of interrater agreement measures", Can. J. Stat. **27**, 3–23.

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

Barry, W., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. (**2003**). "Do rhythm measures tell us anything about language type?", in *Proceedings of the 15th ICPhS 2003, Barcelona*, edited by M. Solé, D. Recasens, and J. Romero, 2693–2696 (Causal Productions Pty Ltd, Barcelona).

Barry, W. and Russo, M. (**2003**). "Measuring rhythm: is it separable from speech rate?", in *Actes des interfaces prosodiques*, edited by A. Mettouchi and G. Ferré, 15–20 (Université Nantes, Nantes).

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

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. S. (**2006**). "Evidence for attractors in English intonation", J. Acoust. Soc. Am. **119**, 4006–4015.

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

Dellwo, V. (**2006**). "Rhythm and speech rate: a variation coefficient for deltac", in *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium, Piliscsaba 2003*, 231–241 (Peter Lang, Frankfurt).

Dellwo, V., Fourcin, A., and Abberton, E. (**2007**). "Rhythmical classification of languages based on voice parameters", in *Proceedings of the International Congress of Phonetic Sciences (ICPhS) XVI, Saarbrücken, 6-10 August, 2007*, 1129–1132.

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

Efron, B. (**1982**). *The Jackknife, the Bootstrap, and other resampling plans*, number 38 in CBMS-NSF Regional Conf. Series in Applied Mathematics (Philadelphia: SIAM).

Ferragne, E. and Pellegrino, F. (**2004**). "A comparative account of the suprasegmental and rhythmic features of British English dialects", in *Actes de Modelisations pour l'Identification des Langues, Paris, 29-30 novembre 2004*, 121–126 (Paris).

Galves, A., Garcia, J., Duarte, D., and Galves, C. (**2002**). "Sonority as a basis for rhythmic class discrimination", in *Speech Prosody 2002, Aix-en-Provence*, 323–326.

Gerhardt, K., Abrams, R., and Oliver, C. (**1990**). "Sound environment of the fetal sheep", Am. J. Obstet. Gynecol. **162**, 282–287.

Grabe, E. and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis", in *Laboratory Phonology, 7*, edited by C. Gussenhoven and N. Warner, 515–46 (Mouton de Gruyter, Berlin, Germany).

Ho, T. K. (**1998**). "The random subspace method for constructing decision forests", IEEE T. Pattern. Anal. **20**, 832–844.

Iversen, J. R., Patel, A. D., and Ohgushi, K. (**2008**). "Perception of rhythmic grouping depends on auditory experience", J. Acoust. Soc. Am. **124**, 2263–2271.

Keane, E. (**2006**). "Rhythmic characteristics of colloquial and formal Tamil", Lang. Speech **49**, 299–332.

Kochanski, G. (**2010**a). "Python package g_classifiers-0.30.1", Software Download, University of Oxford, URL `https://sourceforge.net/projects/speechresearch/files/g_classifiers/g_classifiers-0.30.1/g_classifiers-0.30.1.tar.gz/download`, downloaded on 12 August 2010.

Kochanski, G. (**2010**b). "Python package gmisclib-0.67.9", Software Download, University of Oxford, URL `https://sourceforge.net/projects/speechresearch/files/gmisclib/gmisclib-0.67.9/gmisclib-0.67.9.tar.gz/download`, downloaded on 12 August 2010.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (**2005**). "Loudness predicts prominence: Fundamental frequency lends little", J. Acoust. Soc. Am. **118**, 1038–1054.

Kochanski, G., Loukina, A., Keane, E., Shih, C., and Rosner, B. (**2010**). "Long-range prosody prediction and rhythm", in *Proceedings of Speech Prosody 2010, Chicago*, 100222:1–4.

Kochanski, G. and Orphanidou, C. (**2008**). "What marks the beat of speech?", J. Acoust. Soc. Am. **123**, 2780–2791.

Kochanski, G. and Rosner, B. S. (**2010**). "Bootstrap Markov chain Monte Carlo and optimal solutions for the Law of Categorical Judgment (corrected)", arXiv:1008.1596 [cs.NA], submitted to Behavior Research Methods; preprint available at http://arxiv.org/abs/1008.1596, downloaded 12 Aug 2010.

Komatsu, M. (**2007**). "Reviewing human language identification", in *Speaker Classification II*, 206–228 (Springer-Verlag Berlin Heidelberg).

Ladefoged, P. and Maddieson, I. (**1996**). *The sounds of the world's languages* (Blackwell, Oxford).

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., and Caviness, J. N. (**2009**). "Quantifying speech rhythm abnormalities in the dysarthrias", J. speech. lang. hear. res. **52**, 1334–1352.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (**1953**). "Equations of state calculations by fast computing machines", J. Chem. Phys. **21**, 1087–1091.

Murty, L., Otake, T., and Cutler, A. (**2007**). "Perceptual tests of rhythmic similarity: I. mora rhythm", Lang. Speech **50**, 77–99.

Navrátil, J. (**2001**). "Spoken language recognition - a step towards multilinguality in speech processing", in *IEEE transactions on Speech and Audio Processing*, volume 9, 678–685.

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

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

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

Ramus, F. (**2002**). "Acoustic correlates of linguistic rhythm: perspectives", in *Speech prosody 2002, Aix-en-Provence*, 115–120.

Ramus, F., Dupoux, E., and Mehler, J. (**2003**). "The psychological reality of rhythm classes: Perceptual studies", in *Proceedings of the 15th ICPhS, Barcelona*, 337–340 (Universitat Autònoma de Barcelona).

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

Szakay, A. (**2008**). "Social networks and the perceptual relevance of rhythm: a New Zealand case study", University of Pennsylvania Working papers in linguistics **14**, 148–156.

Tumer, K. and Ghosh, J. (**1996**). "Error correlation and error reduction in ensemble classifiers", Connect. Sci. **8**, 385–404.

Tyler, M. D. and Cutler, A. (**2009**). "Cross-language differences in cue use for speech segmentation", J. Acoust. Soc. Am. **126**, 367–376.

Wagner, P. and Dellwo, V. (**2004**). "Introducing YARD (Yet Another Rhythm Determination) and re-introducing isochrony to rhythm research", in *Speech Prosody 2004, Nara, Japan*, edited by B. Bel and I. Marlien, 227–230.

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

White, L. and Mattys, S. L. (**2007**b). "Rhythmic typology and variation in first and second languages", in *Segmental and prosodic issues in Romance phonology.*, edited by P. Prieto, J. Mascaró, and M.-J. Solé, 237–257 (John Benjamins, Amsterdam).

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

Young, S. J., Evermann, G., Gales, M. J. F., Moore, D. K. G., Odell, J. J., andD. Povey, D. G. O., Valtchev, V., and Woodland, P. C. (**2006**). *The HTK book version 3.4* (Cambridge University Engineering Department, Cambridge, UK), `http://htk.eng.cam.ac.uk/docs/docs.shtml`, downloaded on 20 November 2009.

TABLE I. Rhythm measures used in this study classified by type of intervals, and scope and normalization.

RM	Description	Type of interval	Scope	Normalization	Reference
%V	Percentage of vocalic intervals	Ratio	Global	Yes	Ramus *et al.* (1999)
Vdur/Cdur	Ratio of vowels duration to consonant duration	Ratio	Global	Yes	Barry and Russo (2003)
ΔV	Standard deviation of vocalic intervals	V	Global	No	Ramus *et al.* (1999)
VarcoΔV	ΔV/mean vocalic duration	V	Global	Yes	Dellwo (2006)
VnPVI	Normalised pairwise variability index (PVI) of vocalic intervals	V	Local	Yes	Grabe and Low (2002)
med VnPVI	VnPVI computed using median value	V	Local	Yes	Ferragne and Pellegrino (2004)
ΔC	Standard deviation of consonantal intervals	C	Global	No	Ramus *et al.* (1999)
VarcoΔC	ΔC/mean vocalic duration	C	Global	Yes	Dellwo (2006)
CrPVI	Raw PVI of consonantal intervals	C	Local	No	Grabe and Low (2002)
CnPVI	Normalised PVI of consonantal intervals	C	Local	Yes	Grabe and Low (2002)
med CrPVI	CrPVI computed using median value	C	Local	No	Ferragne and Pellegrino (2004)
PVI-CV	PVI of consonant+vowels groups	CV	Local	No	Barry *et al.* (2003)
VI	Variability index of syllable durations	CV	Local	Yes	Deterding (2001)
YARD	Variability of syllable durations	CV	Local	Yes	Wagner and Dellwo (2004)
nCVPVI	Normalised PVI of consonant+vowel groups	CV	Local	Yes	Asu and Nolan (2005)

TABLE II. The smallest number of RMs and the best performing measures or types of measures for optimal separation of pairs of languages. The best performing RMs depend on the language pair.

Language pair	Min N of RMs	Best RMs
Russian-Mandarin	1	ratio, VarcoΔV, ΔC, %V
English-Mandarin	1	ratio, C
French-Mandarin	1	VarcoΔV, %V
Greek-Mandarin	1	ratio, normalized V and CV
Russian-Greek	3	V and CV
English-Greek	1	VarcoΔV, medCrPVI
English-Russian	1	VarcoΔV
English-French	1	VarcoΔV
French-Greek	1	VnPVI
French-Russian	1	medVnPVI

TABLE III. Percentage of times that SA2a assigned 'C' or 'V' labels to voiceless obstruents, voiced obstruents, vowels and consonants within languages.

	Voiceless obstr.		Voiced obstr.		Sonorant		Vowel	
	'C'	'V'	'C'	'V'	'C'	'V'	'C'	'V'
Russian	88	10	55	45	9	91	7	92
Greek	83	11	54	46	17	83	8	91
French	89	8	56	43	18	80	12	87
Chinese	86	11	41	59	22	77	10	89
English	84	11	75	22	15	83	11	88

LIST OF FIGURES