# Long-Range Prosody Prediction and Rhythm

*Greg Kochanski, Anastassia Loukina, Elinor Keane,*
*Chilin Shih*[†] *and Burton Rosner*

University of Oxford Phonetics Laboratory, Oxford, UK
[†]EALC/Linguistics, University of Illinois, Urbana-Champaign, USA
greg.kochanski@phon.ox.ac.uk

## Abstract

Rhythm is expressed by recurring, hence predictable beat patterns. Poetry in many languages is composed with attention to poetic meters while prose is not. Therefore, one way to investigate speech rhythm is to evaluate how prose reading differs from poetry reading via a quantitative method that measures predictability.

We use linear regression to predict the acoustic properties of segments from the properties of up to 7 preceding segments. This accounts for as much as 41% of the variance of some regressions on our full (prose) corpus and up to 79% in a subcorpus of poetry. While roughly half of the predictive power comes from the segment immediately preceding the target, the predicted variance increases by 6% (for the full/prose corpus) or by 25% (for the poetry sub-corpus) upon extending the predictor to include the seven preceding segments. Therefore, interactions between segments extend well beyond the immediate vicinity. Potentially, these longer-range regressions capture the rhythms of the poetry. This approach could form the foundation of a general method for characterizing the statistical properties of spoken language, especially in reference to prosody and speech rhythm.

**Index Terms**: poetry, rhythm, prosody, syllable, prediction, machine learning

## 1. Introduction

Speech is not isochronous. [1] showed that three syllable-timed languages do not actually have equally long syllables, and three stress-timed languages do not have equally spaced stresses. Yet, there is a common perception, and not just among linguists, that different languages have a different rhythm.

We began work on this problem in [2], where we studied rhythm measures based on durations of vocalic and consonantal intervals. We found the measures had substantial person-to-person and text-to-text variability. Languages could not be identified from the rhythm measures unless averages were conducted over a substantial corpus. This result seems to conflict with the intuition that different language = different rhythm.

To check that intuition, we now cast our net wider, investigating prosodic properties beyond duration and searching for rhythmic predictability over a range of about 4 syllables.[1] This range includes about two prosodic feet or a phrase. Our choice of acoustic properties was inspired by the substantial effects seen in loudness and quasi-duration in our previous work [3].

The essence of rhythm is predictability. For example, in 2/4 music or iambic poetry, one can anticipate the strength of the next beat, given knowledge of the previous ones. To quantify predictability, we use a linear regression of acoustic properties of segments against acoustic properties of preceding segments. Pearson's $r^2$ quantifies the success of the regression.[2] Since we use a time axis defined in segments rather than seconds, we may find predictability where [1] did not.

Previous rhythm measures (see list in [2]) are all time-symmetrical, so they would yield the same value if an utterance were played backwards or forwards. But humans are not time-symmetrical: [4] found that while infants could distinguish between French and Russian when speech was played forwards, they were unable to do so when the speech was played backwards. Since previous rhythm measures cannot make these distinctions and no machine learning system based upon them can do so, there must be rhythmic information that existing measures do not capture.

## 2. Methods

Our corpus consisted of the data collected for [2]; it contains 42 paragraphs read in each of Standard Modern Greek, Parisian French, Southern British English, Standard Russian, and Taiwanese Mandarin.

There were 10 readers for each language (24 readers for English). Readers were native speakers of the appropriate language who (except for the English) had lived in English speaking areas for three years or less. Every speaker also read 4 short poems consisting of 8-12 lines. For English, Russian and Greek we selected poems composed in iambic or trochaic tetrameter. French poems contained 8 syllables in each line. Where possible, we selected children's poetry which had a regular metrical pattern and was likely to be read with a strong rhythm. For Mandarin we selected nursery rhymes readable in trochees and a modern poem readable in an iambic meter. We did not instruct subjects to read the poetry in a rhythmic or expressive style.

We used paragraphs whose syllable count (as predicted from the text[3]) and recording duration were both between the $10^{\text{th}}$ and $90^{\text{th}}$ percentiles of the corresponding distributions for that language. Averaged over languages, the limits were 64.4 to 205.6 predicted syllables and 24.7 seconds to 67.3 seconds.[4]

[1] Prior work has generally not used phonologically defined syllables, but often the lengths of successive vocalic regions, which are likely to form the cores of adjacent syllables. We use the term in this latter, approximate sense.

[2] Recall that Pearson's $r^2$ is the fraction of the data's variance that can be explained by the linear regression.

[3] We used [5] to transcribe French texts. Speech Technology Center Ltd. (St.-Petersburg, Russia) transcribed the Russian for us, and the Institute for Speech and Language Processing (Athens, Greece) transcribed the Greek. Their help is gratefully acknowledged.

[4] The poetry is short, so these limits exclude almost all of it from the "full" corpus; it can be safely thought of as a "prose" corpus.

### 2.1. Segmentation

We designed our segmentation procedure to be strictly language-independent, to enable comparison of results across languages.[5] To do this, we first processed the recordings by computing an acoustic description vector, based on [6]. Then we built a specialised speech recognition system, based on the HTK toolkit, [7] that produced a sequence of **C** (consonantal), **V** (vowel-like) and **S** (silence/pause) segments.

The system was trained on 17 paragraphs of the data that were manually segmented into phones by the authors. Each language had at least one segmented paragraph. In the training data, all vowels and sonorants were mapped into **V** and all other phonemes into **C**. Pauses were manually identified and transcribed as **S**.

The recogniser was a monophone system: the **C** segment had six states and a minimum duration of 20 ms, **V** had six states and a minimum duration of 30 ms, and **S** had 13 states and a minimum duration of 110 ms. Eight of the states had four Gaussian mixtures; the others had just one. All the Gaussians had independent diagonal covariances; nothing was tied. The acoustic feature vector for the recogniser was 41-dimensional, differing primarily from the acoustic description vector in [6, 8] in that the spectral components were not smoothed. Additionally, it included analogues of the "Ldur" and "Fdur" values defined below in §2.2 averaged over the preceding 250 ms.

### 2.2. Predictors

Once the segment boundaries were defined, five acoustic properties were computed for each segment. These were used as the independent variables in the linear regressions. We computed:

1. $\log(\Delta)$, where $\Delta$ is the segment duration.
2. $\mathrm{Ldur} = \sum L \cdot \delta t/(t_0 + \Delta)$ ("loudness"), where $t_0 = 100$ ms, $\delta t = 10$ ms is the time step, and the sum is taken over the segment. This uses the loudness density estimator from [3, 9]. It was designed to reflect the perceived loudness. It incorporates (approximately) the result from [10] that the perceived loudness of tone bursts increases with their duration for durations shorter than ≈200 ms.
3. $\mathrm{Fdur} = \sum L \cdot (A - \bar{A}) \cdot \delta t/(t_0 + \Delta)$ ("frication"), as above, but incorporating the aperiodicity measure from [3]. $\bar{A}$ is the average aperiodicity, averaged over the entire paragraph and weighted by loudness. Fdur measures the strength of frication: voiceless fricatives will give positive values, vowels negative, and silences and certain voiced fricatives will yield zero.
4. $\mathrm{Lskew} = \sum L \cdot (t - t_c) \cdot \delta t/(t_0 + \Delta)$ ("loudness skew"), where $t$ is time and $t_c$ indicates the phoneme center. This describes whether the loudness density is concentrated early (negative) or late (positive) in the segment.
5. $\mathrm{dSdt} = log(\sum(\delta t/D)/\Delta)$ ("Spectral Change"), where $D$ is the running duration measure from [3]. Since $D$ measures the time interval between substantial changes in the speech spectrum, the sum can be thought of as the number of substantial spectral changes in the segment, and the overall value can be interpreted as the rate at which spectral changes happen within that segment.

[5] Human segmentation is probably not language independent because it depends on phonological knowledge which incorporates much language-specific information. In [2] we found that acoustic-based segmentation led to different treatment of stop consonants reflecting differences between languages. This raises a potential question for all rhythm measures derived from manual segmentation: are the language-to-language differences inherent in the speech or are they a property of the segmenter?

### 2.3. Bootstrap Resampling

In addition to computing statistical significance of our linear regressions via standard F-tests, we used a bootstrap resampling technique as a second check, e.g. to guard against non-Gaussian distributions of our acoustic properties. We created 25 bootstrap replications of our data and computed the linear regression for each; this gave 25 sets of regression coefficients. Importantly, the coefficients obtained from bootstrap resampling are nearly independent samples, and their distribution mirrors the statistical uncertainty in the regression, almost as if one ran the experiment multiple times.

We used the bootstrap results to obtain a confidence level for each coefficient. We computed the mean and standard deviation of each coefficient's values and then tested the mean against zero with Student's t-test. Likewise, we used the 25 replications for the mean-squared error before and after the regression to compute 25 values of Pearson's $r^2$, which yielded histograms and confidence limits for $r^2$.

## 3. Results and Discussion

### 3.1. Performance of the Recognizer

We first determined whether the recognizer treated human-identified segments consistently. For this, we computed the fraction of each segment's duration that was recognized by the system as vowel-like or consonant-like. Next, we computed the medians for each phone (e.g. segment identified by human segmentation as [t] (475 occurrences) is typically recognized as 89% **C**, 11% **V**, and 0% **S**). Then, for each instance of that phone, we checked if the **V/C** split falls near (within 15% of) the median. For example, for [t], "near" means a **C** fraction between 74% and 100%. Overall, 75% of all instances were near their corresponding median. For comparison, the consistency score between segmentations done by two professional phoneticians was 80%.

We also examined the correlation between the number of segments predicted by the transcription and the number of segments returned by our recognizer (see Figure 1). This comparison showed a close match between the recognized and predicted number of segments (Pearson's $r^2$=0.96), although 11% fewer segments were systematically recognized than expected. This may be due to connected-speech processes such as consonant lenition and vowel reduction. Also note that we trained the system with sonorants mapped to **V**, so that about 20% of the time a **V** region spanned two syllables.

### 3.2. Predicting Acoustics

We computed linear regressions for 31 combinations of acoustic properties and contexts (see the vertical axis for Figure 2). We predicted the properties of the central segment in the context, so in a "V,C,S" context, we predicted the properties of a **C**, and we selected ones found between **V** and **S**: in other words, a phrase-final consonantal region. For each combination we computed linear regressions involving a constant plus the acoustic properties of up to seven preceding segments. If there were any preceding silences in range not specified in the context, we dropped the datum. That is, we predicted from a strict **C**, **V**, **C**, . . . al-

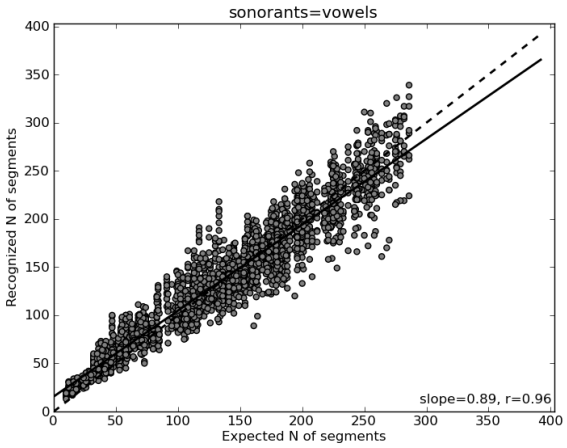


Figure 1: Number of vocalic segments in the expected transcription (horizontal axis) compared to the number produced by the recognizer (vertical axis). The dashed line indicates equality, and the solid line is the best linear fit to the data.

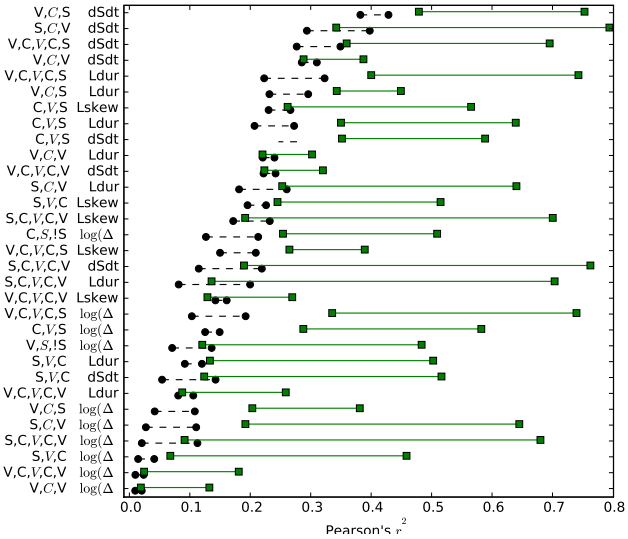


Figure 2: Pearson's $r^2$ (horizontal axis) for prose and poetry for different combinations of context and the target of prediction (vertical axis). The lines connect the value of $r^2$ for 6-parameter linear regressions that use only the immediately preceding segment to predict, to the $r^2$ for 36-parameter models that use data from seven segments before the prediction target. Dashed lines mark regressions on the entire corpus and solid lines mark regressions on the poetry sub-corpus. Lines have markers if the endpoints are significantly different at the 0.001 level.

ternation, broken if and where the context specified a silence. These linear regressions for each of our five languages yielded a total of $7 \cdot 31 \cdot 5 = 1085$ regressions, each evaluated with 25 bootstrap data sets.

Of the 1085 regressions, 97.5 per cent were statistically significant. This is a consequence of the large amount of data: the regressions had 529 to 110133 degrees of freedom (mean 11486). Note that we were not predicting the difference between vowels and consonants: our context selected either the one or the other for each regression. Therefore, the regressions show the variation of properties within those classes of sounds.

### 3.3. Predicting Properties of Prose

For the regressions over the entire corpus, values of Pearson's $r^2$ varied widely: some regressions explain a negligible 2% of the total variance, and others up to 43%.[6] Duration is the least predictable property: on average, $r^2$ was only 8% (min 2%, max 19% over contexts). This is noteworthy because all the published rhythm measures are based only on duration. At the other extreme, the most predictable property is dSdt, with an average $r^2$ roughly three times larger (mean 27%, min 17%, max 41% over contexts).[7] Plausibly, dSdt is related to hyper/hypo-articulation of a segment: more complete articulation might be expected to yield larger changes in the spectrum within a segment and therefore, larger values of dSdt.

Prediction of the properties of phrase-final segments was more effective than for phrase-medial or phrase-initial segments. The mean $r^2$ for phrase-final segments was 26%, versus 19% for -initial and 16% for -medial. Effective prediction of phrase-initial properties was surprising, since they are predicted based on the preceding segments, which comprise a silence and the tail end of the previous phrase. We did not expect silences to be informative nor did we expect much correlation from one phrase to the next. One interpretation of this result is that the pause is planned together with the beginning of a phrase. If they are planned as a unit, then correlations of the pause duration with properties of the phrase-initial segments could occur. Such global planning has been shown for inspired lung volume and pitch as a function of sentence length [11, 12, 13, 14].

The regressions were about equally successful at predicting the properties of vocalic and consonantal regions: the ten best lines in contexts are split half-and-half between predicting **C** and **V**. Figure 2 summarizes the results of linear regressions showing different amounts of prior context, for the entire corpus and also for the "poetry" fraction of it. The data displayed here have 41 to 3977 degrees of freedom, with a mean of 526.

Finally, we note that – averaged over all targets and contexts – all five languages are be nearly equally predictable. In our analysis the range is only from 16% (English) to 19% (Mandarin and Greek). However that ranking seems to be context- and target-dependent, so an analysis looking at different texts or acoustic properties could very well give a different ranking.

### 3.4. Predicting properties in poetry

Overall, poetry is much more predictable than prose ($r^2$ values are roughly twice as large). This is consistent with the intuition that poetry is more 'rhythmical'.

Figure 2 shows a different pattern of $r^2$ between prose and poetry. Overall, phrase-final and phrase-initial segments were more predictable in poetry. Averaging over all targets, $r^2$ for phrase-final was 45% with 38% for -initial and 20% for -medial; all are above the corresponding averages for the full corpus. Considering that the pauses in poetry most commonly occurred between the lines, our algorithms show that the acoustic properties of the first and last segment of each line are highly predictable.[8] At the same time, there was little difference in pre-

[6] Pearson's $r^2$ is averaged over the bootstrap replications.

[7] This large value suggests that predictability is not specific to a single language. It is mathematically impossible for all that $r^2$ to be concentrated within a single language.

[8] When we compute $r^2$, we compare a full model to a simple model

dictability of spectral change (dSdt) of phrase-medial segments. This is consistent with our hypothesis that dSdt may reflect hyper/hypoarticulation and thus in phrase-medial position should not be affected by the differences between prose and poetry.

We also observed that the long-range effects were stronger in poetry than in prose. While in prose mean difference between $r^2$ for the regressions based on 1 and 7 preceding segments was 6%, in poetry this difference was 25%. Given that all poetry in our corpus had regular metrical pattern, this confirms that the long-range effects we observe are likely to be related to such linguistic patterns as feet.

Predictable acoustic properties can have very different interpretations, depending on what they are predicted *from*. For instance, a local effect, depending only on the immediately preceding segment, could be a universal, physiological limitation of muscle motion or motor planning, while a longer range prediction would suggest linguistic patterns and might correspond to feet and rhythm. This is what we observe, and the effect is stronger in poetry.

In poetry, the overall average $r^2$ is also (as for prose) nearly equal across languages: it ranges from 35% (Greek and French) to 39% (Russian). This is somewhat surprising, as descriptions of French and Mandarin poetry are not obviously centered on a stressed/unstressed alternation. There may be a common poetic reading style that applies at least to children's poetry in most languages, even though it may be described differently.

# 4. Conclusions

We investigated the predictability of certain acoustic properties for speech, where the prediction is based on the properties of the preceding 1 through 7 segments. All the acoustic properties used were plausibly related to prosodic properties of the speech.

We found substantial and statistically significant differences in predictability (measured by Pearson's $r^2$ of linear regressions) from one acoustic property to another and from one context to another. Languages, averaged over all contexts and targets have similar values of $r^2$, but different languages may have different patterns of high and low $r^2$ as a function of context and target.

We argue that more rhythmical styles of speech will be more predictable from previous segments, if those previous segments span several prosodic feet. For the full corpus, which is dominated by prose, a consistent, modest increase in $r^2$ occurred as we extended our predictor from the single preceding segment out to a 7-segment long predictor. Almost every combination of target and context that we investigated is influenced by longer-range interactions as well as those from the immediately preceding segment.

Our poetry sub-corpus, was much more predictable than the prose and had a much larger increase in $r^2$ as the predictor extended from 1 to 7 segments. This is consistent with the poetry being read in a much more rhythmic style, as would be expected.

The predictability of a language depends on what is being predicted and the context of the target phones, so we anticipate that there will be at least several different ways to characterise the rhythm of each language. We propose that using linear regressions to predict segmental properties in terms of previous segmental properties could be the foundation of a new set of rhythm measures.

that just contains a single constant. So, this is not just a statement that the final segment is stressed (or unstressed, as the case may be). If the final segment were always stressed, the information would be captured by the constant term and would not be included in $r^2$. A plausible interpretation is that sometimes it is stressed and sometimes not, and the values of $r^2$ correspond to the ability to predict which.

# 5. Acknowledgements

We thank John Coleman, Anne Cutler, and Daniel Hirst for comments and questions. This project is supported by the Economic and Social Research Council (UK) via RES-062-23-1323, and we acknowledge the National Science Foundation (US) for supporting Dr. Shih via IIS-0623805 and IIS-0534133.

# 6. References

[1] P. Roach, "On the distinction between 'stress-timed' and 'syllable-timed' languages," in *Linguistic Controversies*, D. Crystal, Ed. London: Edward Arnold, 1982, pp. 73–79.

[2] A. Loukina, G. Kochanski, C. Shih, E. Keane, and I. Watson, "Rhythm measures with language-independent segmentation," in *Proceedings of Interspeech 2009 : Speech and intelligence*. International Speech Communications Association, 2009, pp. 1531–1534, brighton, UK, 6-10 September.

[3] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *J. Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[4] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amieltison, "A precursor of language acquisition in young infants." *Cognition*, vol. 29, no. 2, pp. 143–178, 1988.

[5] F. Bechet, "Lia_phon: Un systeme complet de phonetisation de textes," *Traitement Automatique des Langues*, vol. 42 (1), no. 1, pp. 47–67, 2001.

[6] G. Kochanski and C. Orphanidou, "Testing the ecological validity of speech," in *Proceedings of the 16th International Congress of Phonetic Sciences*, J. Trouvain and W. J. Barry, Eds., 2007. [Online]. Available: http://kochanski.org/gpk/papers/2007/icphs.pdf

[7] S. J. Young, G. Evermann, M. J. F. Gales, D. K. G. Moore, J. J. Odell, D. G. O. andD. Povey, V. Valtchev, and P. C. Woodland, *The HTK book version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006, http://htk.eng.cam.ac.uk/docs/docs.shtml, checked 11/2009.

[8] L. Baghai-Ravary, G. Kochanski, , and J. Coleman, "Precision of phoneme boundaries derived using Hidden Markov Models," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, ser. ISSN 1990-9772, ISCA, Ed., 2009, pp. 2879–2882, interspeech 2009, Brighton, UK, 7-10 September 2009.

[9] G. Kochanski and C. Orphanidou, "What marks the beat of speech?" *J. Acoustical Society of America*, vol. 123, no. 5, pp. 2780–2791, 2008, URL viewed 7/2008. [Online]. Available: http://kochanski.org/gpk/papers/2006tapping.pdf

[10] R. Plomp and M. A. Bouman, "Relation between hearing threshold and duration for tone pulses," *J. Acoustical Society of America*, vol. 31, no. 6, pp. 749–758, June 1959.

[11] A. L. Winkworth, P. J. Davis, R. D. Adams, and E. Ellis, "Breathing patterns during spontaneous speech," *Journal of Speech and Hearing Research*, vol. 38, no. 1, pp. 124–144, 1995.

[12] D. H. McFarland and A. Smith, "Effects of vocal task and respiratory phase on prephonatory chest-wall movements," *J. Speech and Hearing Research*, vol. 35, no. 5, pp. 971–982, 1992.

[13] D. H. Whalen and J. M. Kinsella-Shaw, "Exploring the relationship of inspiration duration to utterance duration," *Phonetica*, vol. 54, pp. 138–152, 1997.

[14] C. Shih, "A declination model of Mandarin Chinese," in *Intonation: Analysis, Modelling and Technology*, A. Botinis, Ed. Kluwer Academic Publishers, 2000, pp. 243–268.