A deep learning approach to automatic characterisation of rhythm in non-native English speech

Konstantinos Kyriakopoulos, Kate M. Knill, Mark J.F. Gales

ALTA Institute / Engineering Department
Cambridge University
Trumpington St, Cambridge CB2 1PZ, UK
{kk492, kate.knill, mjfg}@eng.cam.ac.uk

Abstract

A speaker's rhythm contributes to the intelligibility of their speech and can be characteristic of their language and accent. For non-native learners of a language, the extent to which they match its natural rhythm is an important predictor of their proficiency. As a learner improves, their rhythm is expected to become less similar to their L1 and more to the L2. Metrics based on the variability of the durations of vocalic and consonantal intervals have been shown to be effective at detecting language and accent. In this paper, pairwise variability (PVI, CCI) and variance (varcoV, varcoC) metrics are first used to predict proficiency and L1 of non-native speakers taking an English spoken exam. A deep learning alternative to generalise these features is then presented, in the form of a tunable duration embedding, based on attention over an RNN over durations. The RNN allows relationships beyond pairwise to be captured, while attention allows sensitivity to the different relative importance of durations. The system is trained end-to-end for proficiency and L1 prediction and compared to the baseline. The values of both sets of features for different proficiency levels are then visualised and compared to native speech in the L1 and the L2.

Index Terms: prosody, rhythm, CALL, speech recognition

1. Introduction

Characterising the prosody of non-native speakers is of increasing interest in the areas of Computer Assisted Language Learning (CALL), automatic assessment and accent detection [1] [2] [3]. An important component of prosody is rhythm, defined as the pattern of phone, syllable and word durations in a person's speech. Different languages have different characteristic natural rhythms, the ability to capture which is a key predictor of the proficiency of a non-native speaker. As a learner's proficiency improves, their rhythm is expected to become less similar to that of their native language and more similar to that of native speakers of the L2.

This paper investigates the extraction of features from real speaker audio to represent rhythm for the purposes of automatic assessment and L1 detection. It is desired for features to be compact, representative and applicable to multiple L1s and tasks. Section 2 explores the features used in the literature to quantify rhythm and motivates the choice of baseline features. Section 3 introduces deep rhythm features, based on attention over an RNN, as a tunable generalisation of the baseline features. Section 4 presents the data and speech recognition system used in the experiments, while Sections 5 and 6 present the results and conclusions.

2. Rhythm Features

Traditionally, the natural rhythm of languages was believed to be governed by a principle known as isochrony, first introduced by Pike [4]. In languages such as French, known as syllable-timed, every syllable takes an equal amount of time to pronounce, while in languages such as English, known as stress-timed, it is the time between the stressed syllables of adjacent words which remains constant. The duration of individual syllables in English is therefore highly variable, depending on where they are relative to the stress of the current and adjacent words. Part of what sounds strange about non-native speech under this theory is a failure to match the stress-timing rhythm of English [5]. This would suggest that the standard deviation of stress-to-stress intervals should be indicative of English proficiency. On the basis of this theory, Honig et al. [6] introduced *isochrony features*:

1. mean and standard deviation of length of time between consecutive stressed syllables
2. mean and standard deviation of length of time between consecutive unstressed syllables
3. ratios of above two means and above two standard deviations

Together (and particularly when combined with tempo features), these features should provide an unbiased metric of the speaker's adherence to the isochrony of English, adjusted for other elements of their voice quality (e.g. how fast they naturally speak).

The main problem with this approach arises from issues with the underlying theory. Firstly, not all varieties of English are stress-timed and those that are are stress-timed to different extents [7]. This could lead to bias based on the variety of native English speech the learner is trying to emulate. In addition, the paradigm of isochrony itself is highly controversial, due to lack of direct empirical evidence of the phenomenon and the failure to classify many languages [8, 9].

The problems with simple isochrony features led Ramus et al. [10] to develop three new features which could be more reliably used to classify languages, based on the properties of adjacent vowels and the intervals between them. Speech is divided into vocalic and consonantal intervals and the following statistics computed:

1. %V: The proportion of time devoted to vocalic intervals in the sentence, disregarding word boundaries
2. ΔV: The standard deviation of the duration of vocalic intervals
3. ΔC: The standard deviation of the duration of consonantal intervals

In a language such as English in which vowels are routinely shortened depending on their position within a word, ΔV and ΔC are very high, while %V is very low (in fact they were respectively the highest and lowest of all languages tested). A low-proficiency non-native speaker with an L1 in which this is not the case is likely to fail to shorten vowels correctly and should therefore fall more closely to their L1 on these three axes. Honig named these features (together with normalised versions of the latter two) *Global Interval Proportions* (GIP) and used them with limited success to predict proficiency [6].

Grabe and Low [11, 12] generalised this concept to develop a more robust metric of rhythm based on the pairwise variability index (PVI), which measures the variability between successive measurements. PVI is applied to the duration of vowels as well as of inter-vocalic intervals. Raw PVI is defined as:

$$\texttt{rPVI} = \frac{1}{m-1}\sum_{k=1}^{m-1}|d_k - d_{k+1}| \tag{1}$$

where d_k is the duration of the kth segment and m is the number of segments. The extraction of rPVI is illustrated in Figure 1.

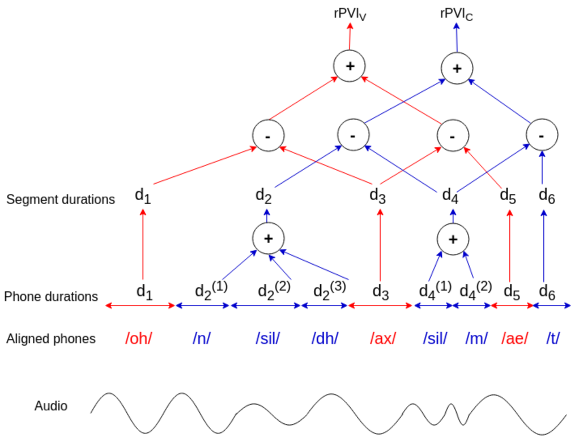

Figure 1: *Illustration of extraction of r-PVI features from sample phrase 'on the mat'*

A normalised version of PVI (nPVI) is also defined as:

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

It was found to improve on rPVI as it adjusts for the speaker's articulation rate and the duration of the particular syllables in question. The authors found that both rPVI and nPVI significantly outperform the Ramus metrics as well as other isochrony metrics at classifying languages based on their rhythmic properties.

Bertinetto et al. [13] modified PVI based on the idea that it is the lengths of individual vowels and consonants, rather than vocalic and consonantal segments, the variation of which is key to the rhythmic properties of languages. They therefore divided the duration of each segment by the number of phones it contained to yield a measure which they term the Control Compensation Index (CCI):

$$\texttt{CCI} = \frac{1}{m-1}\sum_{k=1}^{m-1}\left|\frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}\right| \tag{3}$$

where d_k and n_k are the duration and number of segments of the kth measurement and m is the number of measurements. Languages like English are in their analysis termed 'compensation' languages, in that the sizes of adjacent vowels and adjacent consonants vary to compensate for each other, resulting in them having high CCIs. Speakers of 'control' languages like Italian, try to keep phonemes at a constant length and so have low CCI.

Based on the above work, Honig et al. [6] define six PVI-based features for use in proficiency assessment, namely rPVI, nPVI and CCI for each of vocalic and consonantal segments. Support Vector Machine (SVM) regression is then used to predict human judgments of the acceptability of subjects' rhythm and melody using these and the previously features. The PVI-based features outperform both isochrony and GIP features, but the combination of PVI and GIP performs even better, suggesting that they each contribute different information about the speaker's rhythm.

Based on this analysis, a baseline set of thirteen features is defined consisting of Honig's six features, the three GIP features, mean vocalic and consonant segment durations and the ratio of mean to standard deviation of each of vocalic and consonant segment durations.

These hand-crafted features capture various aspects of rhythm based on theoretical knowledge of the properties of English and other L1s, but are still limited in both their power and their general applicability, due to the host of assumptions they rely on. For this reason, more recent work has focused on generalising PVI to a parametarised function of d_k and d_{k+1}, the parameters of which can be tuned for different tasks and for different mixes of L1s [14]. This approach significantly improves performance on language classification tasks. However, like with PVI and CCI, it is still not possible to capture duration relationships beyond the segment pair level and the summation forced all segments to be given equal weight, when some may actually be more salient to characterising rhythm than other.

Section 3 presents a further generalisation of rhythm features using deep learning, to tackle these two issues and improve tunability.

3. Deep Rhythm Features

In Kyriakopoulos et al. [15], it was seen how recurrent neural layers and attention mechanisms could be used to create a tunable, end-to-end trainable, deep learning alternative to hand-crafted features for characterising pronunciation. The recurrent layers allowed patterns over time to be captured, while the attention mechanisms allowed the relative salience of different time steps to be weighed when compressing the sequences to lower dimensional fixed-length representations. In this section, a similar approach is employed to create a deep alternative to the hand-crafted rhythm features presented in Section 2.

Consider an utterance spoken by a given speaker and divided into V vocalic segments and C intervocalic segments (e.g. the phrase 'on the mat' consists of vocalic segments /oh/, /ax/ and /ae/ and intervocalic segments {/n/,/sil/,/dh/}, /m/ and /t/).

For each of the two types of segments, we define the vector $\boldsymbol{d}_k^{(n)}$, containing the duration, phone identity and other salient information for the kth sub-segment of a given segment n. In PVI, the durations of the sub-segments of each segment are combined by addition, whereas in CCI their mean is computed.

Here, self-attention over the sub-segments is used to capture the relative salience of each, and the result concatenated with the total duration of the segment d_n, to produce the vector $\boldsymbol{d}_n$ representing what we know about the segment n:

$$\boldsymbol{d}_n = \left[\sum_{k=1}^{K^{(n)}} \alpha_k \boldsymbol{d}_k^{(n)}, d_n \right] \tag{4}$$

where

$$\alpha_k = \frac{\exp s(\boldsymbol{d}_k^{(n)}, \boldsymbol{\theta}_{att})}{\sum_{j=1}^{K^{(n)}} \exp s(\boldsymbol{d}_j^{(n)}, \boldsymbol{\theta}_{att})} \tag{5}$$

Next, the sequence of vectors $\boldsymbol{d}_n$ for each of all vocalic and all inter-vocalic segments in each speaker's speech is passed through a bi-directional LSTM to capture dependencies across the whole sequence of durations rather than just pairs of adjacent durations:

$$\boldsymbol{h}_{1:V}^{(V)} = F_{\texttt{LSTM}}(\boldsymbol{d}_{1:V}^{(V)}, \boldsymbol{\theta}_v) \tag{6}$$

$$\boldsymbol{h}_{1:C}^{(C)} = F_{\texttt{LSTM}}(\boldsymbol{d}_{1:C}^{(C)}, \boldsymbol{\theta}_c) \tag{7}$$

Further attention mechanisms project each of the resulting sequences to fixed length vocalic and intervocalic features, to capture the relative salience of each segment to the overall rhythm characterisation:

$$\tilde{\boldsymbol{h}}^{(V)} = \sum_{v=1}^{V} \alpha_v \boldsymbol{h}_v^{(V)} \tag{8}$$

$$\alpha_v = \frac{\exp s(\boldsymbol{h}_v^{(V)}, \boldsymbol{\theta}_{att})}{\sum_{j=1}^{V} \exp s(\boldsymbol{h}_j^{(V)}, \boldsymbol{\theta}_{\texttt{att}})} \tag{9}$$

$$\tilde{\boldsymbol{h}}^{(C)} = \sum_{c=1}^{V} \alpha_c \boldsymbol{h}_c^{(C)} \tag{10}$$

$$\alpha_c = \frac{\exp s(\boldsymbol{h}_c^{(C)}, \boldsymbol{\theta}_{att})}{\sum_{j=1}^{C} \exp s(\boldsymbol{h}_j^{(C)}, \boldsymbol{\theta}_{att})} \tag{11}$$

This system is illustrated in Figure 2. The features $\tilde{\boldsymbol{h}} = [\tilde{\boldsymbol{h}}^{(V)}, \tilde{\boldsymbol{h}}^{(C)}]$ can now be used to represent the speaker's overall rhythm and can be projected through a simple feed forward layer to predict the speaker's grade or accent.

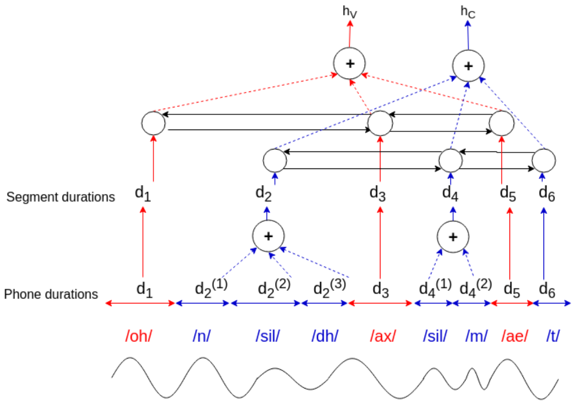

Figure 2: *Illustration of extraction of deep rhythm features from sample phrase 'on the mat'*

The experimental setup for implementing the above and comparing it to the baseline features is explained in Section 4.

4. Experimental Setup

The preceding sections of this paper have described two approaches for extracting rhythm features from the aligned phone sequence of the utterances produced by a candidate and using them to predict their proficiency and native language. The systems are now implemented using TensorFlow and trained and tested on real data.

The data for training and testing are obtained from candidate responses to the spoken component of the Business Language Testing Service (BULATS) for foreign learners of English, provided by Cambridge English Language Assessment. The BULATS speaking test has five sections, all related to business scenarios [16]. Section A consists of short responses to prompted questions. Candidates read eight sentences aloud in Section B. Sections C-E consist of spontaneous responses of several sentences in length to a series of spoken and visual prompts. Candidates are scored on a scale from 0 to 30, based on their overall proficiency, and it is this score that the system is predicting.

The systems are trained on a gender and proficiency level balanced mixed L1 dataset (TRN) consisting of 3376 speakers (first languages Polish, Vietnamese, Arabic, Dutch, French and Thai), scored on their overall proficiency by human graders and evaluated on a held out evaluation set (EVL), consisting of 224 speakers of a similar mix of L1s, gender and proficiency, with scores provided by expert human graders.

The first step before passing the data through the system is recognising the text being spoken and aligning the audio to a sequence of phones. Both these tasks are performed using an automatic speech recogniser (ASR). Due to the incorrect pronunciations, grammar and rhythm, related to the speaker's proficiency level and first language (L1), the accuracy of standard commercial "off-the-shelf" ASR systems is too low for non-native learner English.

Instead, the ASR system from Kyriakopoulos et al. [17] (also described in Van Dalen et al. [18]), which is trained on non-native learners of English, is used. This ASR has an overall word error rate (WER) of 47.5% and a phone error rate (PER) of 33.9%, evaluated against crowd sourced transcriptions of EVL. It should be noted that crowd-sourced transcriptions are themselves often noisy, leading these results to likely under-estimate true ASR performance. This problem of crowd-sourcing noise is mitigated as discussed in Van Dalen et al. [18].

Using the setup described above, experiments are run to evaluate the performance of the systems outlined in Sections 2 and 3 on the data, with the results presented in Section 5.

5. Experimental Results

The shallow baseline features from Section 2 and the deep features introduced in Section 3 are extracted for each speaker in the data presented in Section 4 and used for human assigned proficiency score and native language (L1) prediction tasks (in the later case, trained end-to-end). The results are presented in Tables 1 and 2.

Table 1: *Grade prediction, trained on TRN, tested on EVL*

	PCC	MSE
Baseline	0.778	17.6
Deep	0.784	15.8

Table 2: *6-way L1 detection, trained on TRN, tested on EVL*

	Accuracy
Baseline	56.2%
Deep	73.0%

It is seen that, in both cases, the deep system significantly outperforms the shallow alternatives, supporting the hypothesis that the generalised version of the features is able to capture information about rhythm not present in the shallow features. On the grade prediction task, improvement in MSE is greater than improvement in PCC, which is to be expected since MSE is the target the network is being trained to minimise and given the greater tunability of the deep system.

Table 3 below shows the breakdown of L1 classification accuracy by the speakers score (grouped by CEFR level). Two effects were hypothesised to affect this relationship. First, the rhythm of more proficient speakers would be expected to be more similar to native pronunciation in the L2, meaning it should be harder to distinguish the speaker's L1 as their proficiency increases. On the other hand, the utterances of better speakers are easier for the ASR to understand and so can be expected to have better aligned duration information.

Table 3: *Breakdown of L1 detection accuracy (%) by CEFR level*

	<A1	A1	A2	B1	B2	C
Base.	30.0	58.1	50.4	57.9	55.8	60.0
Deep	40.0	64.3	49.0	53.3	74.5	70.0

It is clear that for the baseline features, the former effect dominates and the accuracy of the alignment information seems to be the limiting factor in characterising rhythm. With the deep features, this effect is less pronounced and the relationship between score and L1 detection accuracy is weaker. This could be consistent with the attention mechanism providing more robustness to low quality duration information, therefore reducing the impact of ASR errors on the end-to-end system's performance.

Finally, Table 4 illustrates the effect of respectively combining the shallow and deep rhythm features with the corresponding shallow and deep pronunciation features introduced in Kyriakopoulos et al. [15]. It is seen that combining the features yields a considerable improvement on the performance of each, confirming that the two sets of features are indeed likely measuring different aspects of speaker proficiency. When the deep features are combined, the feature extractors are being trained to be complimentary to each other, explaining the even greater increase in performance.

Table 4: *Grader PCC for prounciation and rhythm features*

	Shallow	Deep
Pronunciation features only	0.790	0.780
Rhythm features only	0.778	0.784
Pronunciation + Rhythm	0.812	0.818

6. Conclusions

An overview of features used in the literature to characterise rhythm was presented and a set of thirteen baseline features defined. Using a deep neural network architecture, these features were used to predict human-assigned proficiency score and L1 based on spontaneous speech by non-native candidates of a spoken English test.

It was then seen how the baseline features can be replaced by more generalised, tunable deep features, extracted using an attention mechanism over a recurrent neural network. Score and L1 prediction tasks were repeated using these new features, with end-to-end training. It was demonstrated that this method yields marked improvements in performance for both score and L1 prediction. Both the baseline and deep features were found to be complementary to phone distance pronunciation features, confirming that each set of features is capturing a different element of speaker proficiency.

For both the shallow and deep methods, the accuracy of the speech recognition and forced alignment systems were important bottlenecks to performance, but the deep method seemed to be more robust to ASR errors.

7. Acknowledgements

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge Assessment, Cambridge University, for supporting this research and providing access to the BULATS data.

8. References

[1] S. M. Witt, “Automatic error detection in pronunciation training: Where we are and where we need to go,” *Proc. International Symposium on automatic detection on errors in pronunciation training*, vol. 6, 2012.

[2] Q.-T. Truong, T. Kato, and S. Yamamoto, “Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours,” *Proc. Interspeech 2018*, pp. 2186–2190, 2018.

[3] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[4] K. L. Pike, *The intonation of American English*. ERIC, 1945.

[5] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: University Press, 1967, p.97.

[6] F. Hönig, A. Batliner, and E. Nöth, “Automatic assessment of non-native prosody annotation, modelling and evaluation,” in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

[7] D. Deterding, “The measurement of rhythm: A comparison of Singapore and British English,” 2001.

[8] R. M. Dauer, “Stress-timing and syllable-timing reanalyzed,” *Journal of phonetics*, 1983.

[9] A. Arvaniti, “Rhythm, timing and the timing of rhythm,” *Phonetica*, vol. 66, no. 1-2, pp. 46–63, 2009.

[10] F. Ramus, M. Nespor, and J. Mehler, “Correlates of linguistic rhythm in the speech signal,” *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[11] E. L. Low and E. Grabe, “Prosodic patterns in Singapore English,” in *Proceedings of the International Congress of Phonetic Sciences, Stockholm*, vol. 3, 1995, pp. 636–639.

[12] E. Grabe and E. L. Low, “Durational variability in speech and the rhythm class hypothesis,” *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[13] P. M. Bertinetto and C. Bertini, “On modeling the rhythm of natural languages,” in *Proceedings of the Fourth International Conference on Speech Prosody*, 2008.

[14] S. Gharsellaoui, S. A. Selouani, W. Cichocki, Y. Alotaibi, and A. O. Dahmane, “Application of the pairwise variability index of speech rhythm with particle swarm optimization to the classification of native and non-native accents,” *Computer Speech & Language*, vol. 48, pp. 67–79, 2018.

[15] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, “A deep learning approach to assessing non-native pronunciation of english using phone distances,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018, 2018, pp. 1626–1630.

[16] L. Chambers and K. Ingham, “The BULATS online speaking test,” *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[17] K. Kyriakopoulos, M. Gales, and K. Knill, “Automatic characterisation of the pronunciation of non-native English speakers using phone distance features,” in *Proceedings of Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

[18] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, “Improving multiple-crowd-sourced transcriptions using a speech recogniser,” in *Proc. of ICASSSP*, Apr 2015.