DISTANT RHYTHMS: COMPUTING FLUENCY

Xuewei Lin and Dafydd Gibbon

Jinan University, Guangzhou and Bielefeld University
Corresponding author: gibbon@uni-bielefeld.de

ABSTRACT

The phonetic component of L2 fluency assessment has been treated in many previous studies with a variety of annotation-based time-domain measurements (syllable, word and phrase rate, speech-pause ratio, filled-unfilled pauses, pause count, mean run duration) and rarely with frequency-domain methods. A new experimental frequency-domain fluency comparison approach is proposed, using automatic detection of low-frequency spectral features and hierarchical clustering. The method is used to compare L2 English readings by speakers of Mandarin (Pǔtōnghuà) at different fluency levels, with the IPA benchmark text "The North Wind and the Sun" as prompt. The automatic method successfully separated intermediate level L2 English readers from an advanced L2 reader and L1 English readers in a pilot study, showing the novel method to be feasible. A subsidiary hypothesis that a specific text format with stressed vowels marked in bold face leads to more consistent and native-like L2 reading was also successfully tested. Tests with larger databases are planned. The code is freely available.

Keywords: fluency, accentedness, L2 assessment, spectral features, rhythm

1. FLUENCY; THE FREQUENCY DOMAIN

Assessment of phonetic factors in L2 speech fluency includes both time-domain measures of partially automatised production skills, such as types of syllable rate and pruning, mean interpausal (‘run’) duration, filled and unfilled pause ratios, and also subjective judgments of ‘accentedness’, which are “operationalized using impressionistic judgments of how far L2 speakers’ pronunciation diverges from a native speaker target” [20]. Both approaches have subjective elements which involve qualitative decisions on accentedness and on the types of unit to be measured: syllables differ in type within and between languages, discourse ‘hedges’ may fail to be counted as non-fluency, reliability of speech timing segmentation varies. Rhythm assessment (as opposed to teaching methods) is not strongly represented, and prosody tends to be reduced to intonational pitch patterning [2], although structural and discourse-level patterns are well-understood [6].

Phonetic, and particularly prosodic aspects of fluency are sometimes neglected in assessment procedures, but they have consequences for fluency in semantics and comprehension: accentedness, including rhythm patterning, can relate to social prejudice [7]. This underlines the need for complementing traditional methods with reliable (reproducible) and ‘objective’ (validated automatic) procedures [10]. The time-domain phonetic parameters are well-known and form a single complex: *rhythmicity* [25, 4, 24]. But stress patterns, and the multiple rhythms which are relevant for intermediate and advanced speakers, tend to be missed by traditional time-domain measurements.

The present exploratory study applies a novel acoustic phonetic rhythm assessment method with automatic frequency-domain analysis of low-frequency (LF) rhythms to a small but well-defined data set and measures divergence among intermediate standard L2 English readers and between these and highly fluent readers. A secondary aim is to test whether text highlighting supports rhythm consistency in reading. Time T and frequency f are closely related, ($f = 1/T$), but spectral analysis and visualisation offer added heuristic and theoretical value [21, 8].

In [18, 12], frequency-domain fluency parameters were measured in the 200 Hz to 4 kHz frequency band, with wavelet transform and principal component analysis. The present approach analyses *salient spectral magnitude peaks* in the low frequency (LF) spectral range, which are interpreted as *rhythm formants*. Rhythm formants from 10 Hz to 1 Hz mark rhythmic grammar units such as syllables and words, while below 1 Hz they mark rhythmic discourse units (corresponding to time domains between 0.1 s and 1 s, and longer than 1 s, respectively). Rhythm differences are modelled as distances between rhythm formant patterns, using standard distance metrics.

Reading fluency was selected partly to constrain the set of variables for the pilot study, partly because reading aloud is an essential skill in many professional activities, and for information and entertainment. Reading fluency has been defined in terms of accuracy, tempo and expression [13, 14], or as reading rapidly, smoothly, effortlessly, automatically (with little conscious attention to subskills of decoding and word recognition) [17]; for general and multimodal reading contexts cf. also [11, 15]. Rhythm relates particularly to tempo, expression and smoothness.

Section 2 describes the data, in Section 3 gives an overview of the method, and in Section 4 the results are discussed. Section 5 summarises conclusions, problems and future work.

2. DATA

2.1. Readers

The investigation uses a small structured dataset. All readings are from the IPA English benchmark text, "The North Wind and the Sun", to facilitate reproducibility. The readers have different English reading proficiencies, ages and genders:

1. British English L1: *S1*, researcher, conservative RP [1], ID NW048-S2; *S2*, researcher, modified RP, recorded for this study, ID NWAS-048-S1.
2. Advanced L2 English: *S3*, female L1 Mandarin speaking university English teacher, advanced L2 proficiency, recorded for this study. 3 recordings from each of 5 text formats, IDs NWAS_A_01 ... NWAS_D_03.
3. Intermediate L2 English: 10 female and 10 male Mandarin speaking students [26], not language majors (not students of *S3*); 1 reading each, diverse gender-prefixed IDs.

Clustering by proficiency is predicted.

2.2. Text prompt formats

In the social media there are countless videos by both professionals and lay influencers with advice on L2 English pronunciation. The use of different text presentation formats with *S3* was motivated by an informal recommendation made in a video by a popular anonymous Chinese English teaching influencer [19]: *Mark all vowels in the text*. This marked-vowel proposal is original, but from a phonetic point of view suggests that marking all vowels would encourage a Chinese-like syllabic rhythm.

The alternative hypothesis proposed here is: *Mark all lexically stressed vowels in the text*. Highlighting only lexically stressed vowels could encourage more consistent stress-oriented English rhythm, since stress position is a major factor in English speech timing. Stress awareness is recognised as a necessary component of pronunciation training for Chinese speakers [4] [5] and requires appropriate assessment methods. Text formats and font effects are already known to affect reading speed [22] and reading by dyslexics [3].

S1, *S2* and the students used unmodified texts with the story as a single paragraph. For *S3*, different text formats were used; *S3* knows the video vowel marking suggestion (*E* below), but was not informed of the purpose of the other format options. The text prompt formats, including *A*, *B* and *C* as distractors, are:

A: Normal text (text as one paragraph).
B: Each sentence on a new line.
C: Clauses and longer phrases on new lines.
D: Normal text, with stressed vowels marked.
E: All vowels marked (but not ‘silent’ vowels).

The texts were read in this order, so rehearsal effects cannot be discounted. The text for the *D*-readings was marked with stressed vowels, following the simplifying principle that all lexical ‘content’ words are stressed, rather than taking rhetoric or information structure into account:

Text D. The North Wind and the Sun
The N**o**rth W**i**nd and the S**u**n were disp**u**ting which was the str**o**nger, when a tr**a**veller came al**o**ng wr**a**pped in a w**a**rm cl**oa**k. They agr**ee**d that the **o**ne who f**i**rst succ**ee**ded in m**a**king the tr**a**veller t**a**ke his cl**oa**k off should be cons**i**dered str**o**nger than the **o**ther. Th**e**n the N**o**rth W**i**nd bl**ew** as h**a**rd as he c**ou**ld, but the m**o**re he bl**ew** the m**o**re cl**o**sely did the tr**a**veller f**o**ld his cl**oa**k ar**ou**nd him; and at l**a**st the N**o**rth W**i**nd gave **u**p the att**e**mpt. Then the S**u**n sh**o**ne out w**a**rmly, and imm**e**diately the tr**a**veller t**oo**k off his cl**oa**k. And s**o** the N**o**rth W**i**nd was obl**i**ged to conf**e**ss that the S**u**n was the str**o**nger of the t**wo**.

3. METHOD

3.1. Demodulation

Figure 1 illustrates the analysis procedure using the first reading (A1) by *S3*; the original sampling rate of 48 kHz was retained. The top panel shows the *amplitude modulated (AM) waveform*. The superimposed outline shows the demodulated low-pass filtered *absolute amplitude envelope*, the LF information signal, which is an approximate correlate of the phonotactic sonority curve, and whose magnitude and duration variations are the main acoustic factors underlying speech rhythms. The mid panel of Figure 1 shows the LF spectrogram of the reading. The bottom panel shows the LF spectrum with the whole recording as FFT window. Frequency modulation (FM), i.e. the F0 track, is also relevant for fluency and accentedness but together with, the modulation-theoretic background, is dealt with elsewhere [8].

The method is related to the algorithms of music identification apps like Shazam [23].

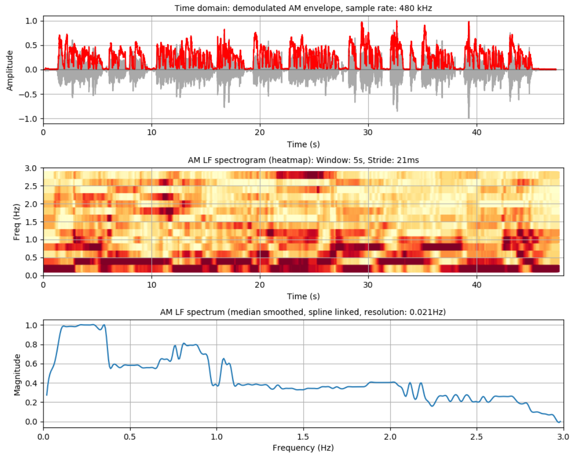

Figure 1: Demodulation and spectral analysis of reading A1 by *S3*. Top: waveform with overlaid amplitude envelope outline. Mid: LF spectrogram. Bottom: LF spectrum.

3.2. Spectral analysis and comparison

The *LF spectrogram* (Figure 1 mid panel) shows rhythm formants as dark horizontal bars: rhythms have minimum durations [16] and constant frequency. The spectrogram is a sequence of spectral slices with 5 s windows (for LF analysis, resolution about 0.21 Hz) with 21 ms steps (for higher time resolution than 5 s). An *LF spectrum* is created by applying an FFT to a window consisting of the entire recording [21, 8] (Fig. 1 bottom panel).

Figure 1 bottom panel shows the *LF spectrum* (resolution 0.021 Hz) with multiple formants: 0.1...0.4 Hz (10...2.5 s beat period), 0.7...0.9 Hz (1.4...1.1 s beat period) and 1.8...2.8 Hz (0.6...0.4 s beat period), which relate to longer sentences, phrases and word constituents.

For comparing the recordings, two sets of vectors are extracted: one from the spectrogram, consisting of the *highest magnitude frequency in each spectral slice*, and one from the spectrum, consisting of the *10 highest magnitude frequencies in the spectrum.*

The vectors in each set are *compared pairwise* using a standard metric, the sum of absolute differences (also known as Manhattan, Cityblock, Taxicab or Mannheim Distance), to indicate divergence from each other and the distances are *hierarchically clustered* with the Nearest Point (Single) linkage algorithm (empirically selected from a range of distance metrics and linkage algorithms). The outputs are agglomerative hierarchical clusterings, visualised as dendrograms, in which the horizontal branch length approximately indicates distances between clusters.

The analysis and plotting procedures are implemented as open source Python prototypes using the NumPy, SciPy, MatPlotLib and Tkinter libraries [9].

4. RESULTS

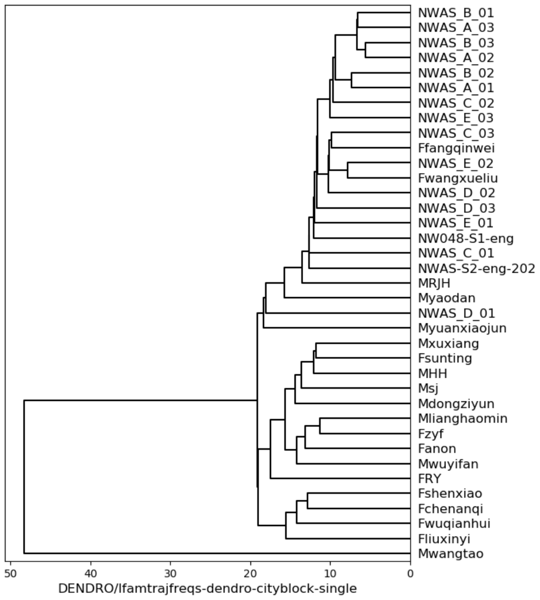

Figure 2: Dendrogram of maximum magnitude frequencies in the LF spectrogram.

The main aim of the experiment is to determine whether spectral features can be used to distinguish between different L2 English proficiency grades, with the subsidiary aim of examining the effect of text formats on consistency of reading.

First results are shown in Figure 2 for time-domain maximum magnitude vectors from the three-dimensional LF spectrogram ($time \times frequency \times magnitude$). The *LF spectrogram-based clusters* in Figure 2 involve not only frequencies but also frequency variation over time, and show a relatively clear distinction, with a few student results intermingled with the native speaker results. Evidently student gender is not a distinguishing feature (cf. M and F name prefixes). However, the fine-grained LF spectrogram vectors may be too high-dimensional for present purposes by over-emphasising high-resolution frequency

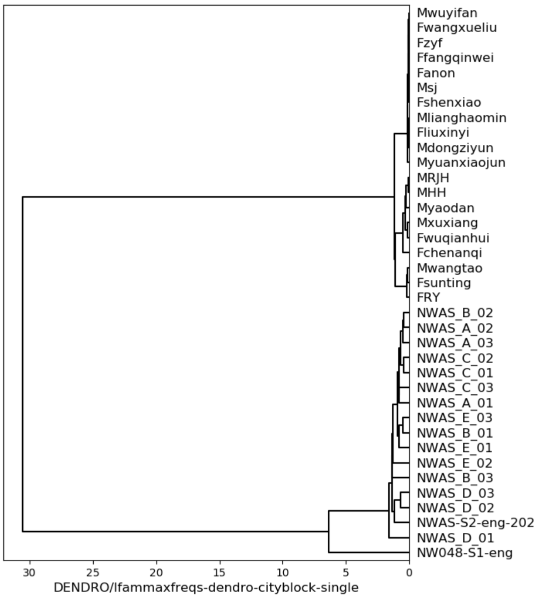

Figure 3: Dendrogram of maximum magnitude frequency similarities.

variation over time in the main frequency zone at the expense of identifying the relevant spectral frequency zones and thus the multiple rhythms.

Figure 3 shows the result for frequency-domain vectors of the 10 highest magnitude frequencies in the *LF spectrum-based clusters* for each reading. The LF spectrum excludes time information, and thus compresses rhythm formants from the entire recording into two dimensions rather than three. Further comparison using Canberra (Normalised Manhattan) and Euclidean distance metrics as a cross-check yields similar results. Pearson Distance provides none of the predicted clusterings: distance between feature vectors seems more important than vector shape difference.

Abstracting away from the time-domain, and considering only the *LF spectrum-based clusters* in the frequency-domain, a clearer picture indeed emerges in Figure 3. There is a partition between the student recordings and the others, and a further partition between *S1* (L1 conservative RP) and the cluster with *S2* (L1 modified RP) and *S3* (advanced L2). With regard to the presentation formats for *S3*, the 3 *D*-format readings cluster together. The student readings cluster without regard to gender.

5. DISCUSSION

The clustering shown in Figure 3 indicates that the first partition separates the intermediate level readers from the highly proficient readers. The 3 bolded-vowel format *D*-readings cluster together, indicating consistent reading by *S3* in this format. Also, clustering of *D*-readings with *S2*, and the next-level cluster with *S1* appear to show native-like low accentedness (high fluency or naturalness) for *S3*.

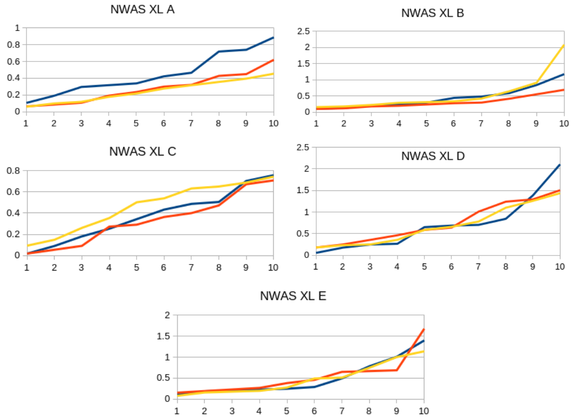

Figure 4: Visualisation of vectors with 10 highest magnitude spectral frequencies: $vectorindex \times frequency$.

Figure 4 shows the 15 extracted spectral peak vectors from the *S3* readings, to enable judgments of subjective plausibility of the distance metric criterion by means of visual inspection. All 3 *D*-readings appear close, while other categories tend to be close pairwise but not overall. A plausible interpretation of this result for the *D*-format style is that bolded stressed vowels constrain consistent conformance with English stress-based timing. It is tentatively claimed, therefore, that in reading aloud (1) spectral features help to determine relative fluency, (2) stressed vowel marking can have a positive effect on consistency of L2 speech timing.

6. CONCLUSION

The present study applied a new automatic acoustic phonetic method to a small but crucial aspect of L2 prosodic fluency which is particularly important for speakers whose L1 and L2 rhythms are radically different, for example Mandarin, which is generally analysed as more syllable-timed, and English, which is generally analysed as more stress-timed.

The results of the present exploratory study clearly need to be calibrated with other methods and scaled up with more data. Analysis of a large classroom data set (150 students) is in progress, and evaluation of results against expert rater judgments.

Further use cases for the method are potentially in components of automatic fluency (self-)assessment and (self-)training apps for medium to advanced L2 teaching, in which further refinement of the method to include more detailed identification of spectral components and adaptation to other language pairs will be required.

7. REFERENCES

[1] D. Abercombie, *The North Wind and the Sun*. University of Edinburgh. School of Philosophy, Psychology, and Language Sciences. Department of Linguistics and English Language, 2013, 1951-1978 [sound]. [Online]. Available: https://doi.org/10.7488/ds/157

[2] J. Barnes and S. Shattuck-Hufnagel, Eds., *Prosodic Theory and Practice*. Cambridge MA: MIT Press, 2022.

[3] BDAtechnology, "Typefaces for dyslexia," 2015, https://bdanewtechnologies.wordpress.com/.

[4] Y. Chen, "A study of rhythmic categorization of stress recurrence in learners' English reading aloud," *Foreign Languages and Foreign Language Teaching*, vol. 3, no. 3, 2015, metadata translated from Chinese.

[5] K. Cheng and Y. Deng, "Preliminary research on relationships between stress awareness and English proficiency among Chinese learners: A case study of Southwest Jiaotong University," *Journal of Southwest Jiaotong University (Social Sciences)*, vol. 16, no. 6, pp. 21–27, 2015.

[6] E. Couper-Kuhlen, *English Speech Rhythm: Form and Function in Everyday Verbal Interaction*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2018.

[7] M. Dehghani, P. Khooshabeh, A. Nazarian, and J. Gratch, "The subtlety of sound: Accent as a marker for culture," *Journal of Language and Social Psychology*, pp. 1–20, 2014.

[8] D. Gibbon, "The rhythms of rhythm," *Journal of the International Phonetic Association, First View*, pp. 1–33, 2021.

[9] ——, "RFAGUI," 2023, Open Source. [Online]. Available: https://github.com/dafyddg/RFAGUI/

[10] D. Gibbon, R. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for spoken Language Systems*. Berlin: Mouton de Gruyter, 1997.

[11] R. F. Hudson, H. Lane, and P. C. Pullen, "Reading fluency assessment and instruction: What, Why and How?" *The Reading Teacher*, vol. 58, no. 8, pp. 702–714, 2005.

[12] H. Kallio, *The Prosody Underlying Spoken Language Proficiency: Cross-lingual investigation of non-native fluency and syllable prominence*. Helsinki: University of Helsinki, 2022.

[13] D. Konza, *Research into practice. Understanding the reading process*. Department of Education and Children's Services, Government of South Australia, 2011, vol. 1.

[14] ——, "Teaching reading: Why the 'Fab Five' should be the 'Big Six'," *Australian Journal of Teacher Education (Online)*, vol. 39, no. 12, 2014.

[15] S. Marzban and Gyöngyi Fábián, "Second language learners' reading strategies. the case of intersemiotic relations," *Argumentum*, vol. 18, pp. 392–409, 2022.

[16] S. Nakamura and Y. Sagisaka, "A requirement of texts for evaluation of rhythm in English speech by learners," in *17th International Congress of Phonetic Sciences*, I. P. Association, Ed. Hong Kong: International Phonetics Association, 2011, pp. 1438–1441.

[17] C. Singleton, *Intervention for dyslexia: A review of published evidence on the impact of specialist dyslexia teaching*. Steering Committee for the "No to Failure" project, Department for Children, Schools and Families, UK, 2009.

[18] A. Suni, H. Kallio, Štefan Benuš, and J. Šimko, "Characterizing second language fluency with global wavelet spectrum," in *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne: International Phonetic Association, 2019, pp. 1947–1951.

[19] A. C. Teacher, "Text marking recommendation," https://www.bilibili.com/video/BV1D14y1s7cx/, 2022, social media video in Chinese.

[20] R. I. Thomson, "Fluency," in *The Handbook of Pronunciation*, M. Reed and J. M. Levis, Eds. Hoboken NJ: Wiley, 2015, pp. 209–226.

[21] S. Tilsen and K. Johnson, "Low-frequency Fourier analysis of speech rhythm," *Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 34–39, 2008.

[22] S. Wallace, Z. Bylinskii, J. Dobres, B. Kerr, S. Berlow, R. Treitman, N. Kumawat, K. Arpin, D. B. Miller, J. Huang, and B. D. Sawyer, "Towards individuated reading experiences: Different fonts increase reading speed for different individuals," *ACM Transactions on Computer-Human Interaction*, vol. 29, no. 4, pp. 1–56, 2022.

[23] A. L.-C. Wang, "An industrial strength audio search algorithm," in *Proceedings of ISMIR 2003, 4th International Conference on Music Information Retrieval*, Baltimore MA, 2003.

[24] L. White and Z. Malisz, "Speech rhythm and timing," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, 2020, pp. 167–182.

[25] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.

[26] J. Yu, "Timing analysis with the help of SPPAS and TGA tools," in *Proceedings of the Tools and Resources for the Analysis of Speech Prosody Workshop*. Aix-en-Provence: Laboratoire Parole et Langage, 2013.