CHARACTERIZING SECOND LANGUAGE FLUENCY WITH GLOBAL WAVELET SPECTRUM

Antti Suni[1], Heini Kallio[1 2], Štefan Beňuš[2 3], Juraj Šimko[1]

University of Helsinki, Finland[1], Slovak Academy of Sciences, Bratislava, Slovakia[2],
Constantine the Philosopher University in Nitra, Slovakia[3]

ABSTRACT

Fluency is a commonly used descriptor of second language (L2) speaking skills. It can be partly conceptualized in terms of appropriate temporal regularity at all levels of prosodic hierarchy, from syllable rate to phrasing. Lack of regularity arising from hesitations, unplanned breaks and repetitions is often present in L2 speech, and negatively affects perceived proficiency.

In this paper, we propose using wavelet spectrum of speech envelope for capturing temporal regularities in L2 speech at various time scales. Analyzing read speech of Slovak L1 and English L1 and L2 speakers, we demonstrate that both syllable and phrase level regularities are distinctly present in the spectra. Also, we show that principal components of the L2 spectra correlate significantly with fluency related assessments.

Keywords: L2 fluency, wavelets, global wavelet spectrum, amplitude modulation

1. INTRODUCTION

Speaking fluently is often the ultimate goal of mastering a second or foreign language (L2). Fluency is also one of the most commonly used criteria in the assessment of L2 skills [13, 15, 3]. This applies both for human assessments as well as for automatic methods of scoring L2 speaking performance [26] and several studies show that automated signal-based scoring correlates well with human assessments [2, 25]. However, there is no consensus of the exact meaning of fluency, and the definitions vary from holistic concepts of overall proficiency [7] to more elaborate phenomena in speech [19].

Breakdown fluency, speed fluency and repair fluency have been identified as major aspects of oral fluency [19]. These aspects can be described as temporal, and many studies on L2 speech focus on measuring one or more of these temporal features [2, 5, 14, 23, 1, 11]. L2 learners tend to speak at slower speech rates than do native speakers [18, 8], and speaking rates appear to become more native-like with overall increase of proficiency. Also stress timing is shown to improve with more experience on the target language [23]. Challenges in the production of stress patterns [9, 17] as well as the presence of various disfluencies in L2 learners' speech such as pauses, hesitations, and repetitions can slow down the speech and cause *temporal irregularities* and irrelevant phrasing, which, in turn, affects the perception of fluency.

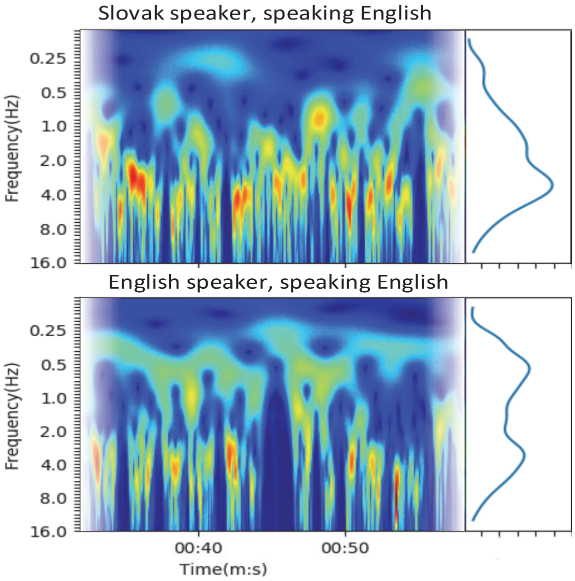
In this study, we examine the concept of fluency as regularity in read speech by analyzing Slovak L1, English L1, and English L2 productions. Namely, we ask if the regularities can be uncovered via frequency analysis; on what frequencies or prosodic scales do these regularities occur; do the L1 and L2 speakers differ; and does the fluency of L2 speakers correlate with global spectral properties.

Time-scale representation based on Continuous Wavelet Analysis (CWT) presents a natural way of depicting a hierarchical structure of a complex signal like speech amplitude envelope. Wide range of time scales, from milliseconds to several seconds in case of prosody, can be analyzed and visualized uniformly due to scale-adaptive windowing [22]. Suitable time vs. frequency resolution compromise can be established even for weakly periodic prosodic signals via selection of the mother wavelet shape.

CWT has previously been applied for detection and quantifying prosodic events regarding word and syllable prominence [12, 20] and phrase boundaries [20, 21, 6]. In comparison, observing the differences between CWT scalograms of L1 and L2 speech (see Fig. 1) led us here to study the *global* aspects on the speech signal by extracting a CWT power spectrum from amplitude envelope. We demonstrate that this completely signal-based representation provides an interesting holistic view of prosody, and that significant proportion of the variance in the spectra can be explained by assessed levels of fluency as well as its other traditional proxies, pausing and speech rate.

While our premises differ, this global approach bears similarities to studies of speech rhythm [16], intelligibility [10] and language comparison [24] using amplitude modulation spectrum, with frequency components of the amplitude envelope obtained using Fourier transform.

Figure 1: Scalograms of L1 and L2 English



2. DATA AND METHODS

2.1. Speech data and assessments

61 speakers of Slovak were recorded in a soundproof studio while reading aloud a story first in Slovak and then in English. The same English text from seven native English speakers (five British, two American) were recorded for reference.

Three Slovak professional teachers of English assessed the fluency of L2 speech samples using a 1–7 Likert scale between “Not fluent at all” (1) and “Very fluent” (7). In addition, ratings for pausing and speech rate were given from “Too many and/or too long pauses” (1) to “Too little and/or short pauses” (7) and “Too slow” (1) to “Too fast” (7) to see, how the perception of these common temporal fluency features are related to actual fluency assessments. Also, overall proficiency was rated on the scale of “Not proficient at all” (1) to “Very proficient” (7).

For the present study, we z-scored the ratings for each individual assessor and used the average ratings across the three assessors.

2.2. CWT analysis

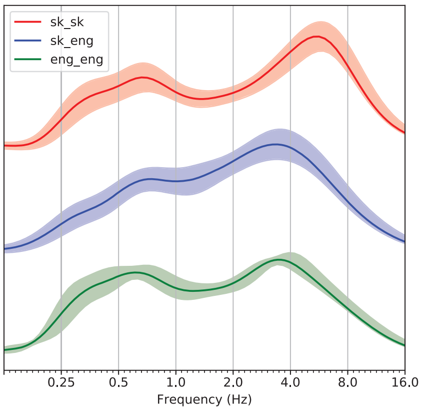
The heat maps in Fig. 1 show examples of wavelet scalograms of two speakers reading the story in English, by a native English speaker (bottom) and by a native Slovak speaker (top).

The analyzed signals are amplitude envelopes of the speech signals extracted with Hilbert transform from 200–4000 Hz band. The envelopes were subsequently low-pass filtered and downsampled to 200 Hz. The dynamic range of the envelopes was then compressed by taking the cube root, yielding a perceptually more plausible signal. Finally, in order to remove the effect of variation in loudness between speakers, the envelopes were normalized to zero mean and unit variance. Wavelet transform with Morlet mother wavelet ($\sigma = 3$) was used covering 7 octaves from 0.125 Hz to 16 Hz, with 10 scales per octave. The red areas in the heatmap in Fig. 1) correspond to high frequency response (peaks) and blue areas to low (negative) response (valleys) .

The global wavelet power spectra, representing the average energy distribution of the envelope in frequency domain, are depicted to the right of each scalogram in Fig. 1. These were calculated from the magnitude scalograms by averaging the responses at each frequency scale over the entire duration of the story [22].

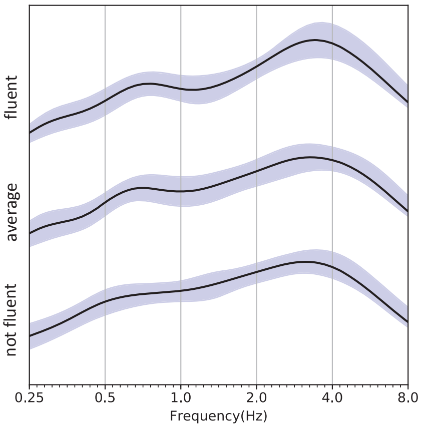
3. RESULTS

Figure 2: Group mean spectra.



The solid lines in Fig. 2 show (from top to bottom) mean global spectra for Slovak speakers reading Slovak story, the same speakers reading the narrative in English, and native English speakers reading the (same) English story. The shaded areas contain values less than standard deviation away from the mean. The mean global spectra for the English narrative have a peak at the frequency just below 4 Hz; for Slovak language story there is an analogous peak at somewhat higher frequency of just over 5 Hz. These peaks presumably correspond to the en-

Figure 3: L2 spectra grouped by assessed fluency.



ergy fluctuations associated with the syllables.

In addition, all mean global spectra exhibit another peak for frequency of approximately 0.6 Hz, likely associated with rhythmic component corresponding to a short phrase. Note, that this peak is less pronounced for English narrative read by the L2 speakers (blue curve in the middle in Fig. 2) compared to the native L1 productions in both languages.

Fig. 3 plots mean global spectra and standard deviation bands for the Slovaks speaking English divided by mean z-scored FLUENCY assessments. The mean assessments were divided to three equally sized groups using 33rd and 66th percentiles, dividing the recordings to, roughly, fluent, average and not-fluent renderings.

The mean global spectra have the peak ≈ 4 Hz. For more fluent speakers, the peak is more pronounced and slightly shifted towards higher frequencies. Also, the low frequency peak at ≈ 0.6 Hz is markedly less prominent for the "not fluent" group compared to the more fluent readings.

To assess these observations quantitatively, we computed Principal Component Analysis (PCA) over the space of all global spectra analyzed in this work, yielding the primary shaping effects explaining most variation among the spectra. The PCA anaysis was calculated on zero centered and scaled (for each frequency) global spectra. The first component explains about 43 %, the first four over 90 % and the first seven components over 98 % of variance.

Linear regressions of the first seven principal components against the mean ratings were used to find out which of these components best correspond to L2 assessments. For each rating we derived the minimal linear model with the rating as the dependent variable and the principal components as the independent variables (a standard procedure of starting with a full model containing all seven components and iteratively pruning the "non-significant" independent variables was used).

Table 1: The fits of PCs againts the ratings.

		effect	t-val	p-val	
FLUENCY	PC1	0.35	3.74	< 0.001	***
	PC4	0.21	2.28	0.026	*
	PC7	0.23	2.48	0.016	*
		Adjusted R-squared: 0.27			
		effect	t-val	p-val	
PAUSING	PC1	0.36	4.22	< 0.001	***
	PC2	0.18	2.14	0.036	*
		Adjusted R-squared: 0.25			
		effect	t-val	p-val	
RATE	PC1	0.36	4.00	< 0.001	***
	PC3	0.45	5.00	< 0.001	***
		Adjusted R-squared: 0.39			

Tab. 1 summarizes the resulting minimal models. (The minimal model for PROFICIENCY ratings was also fitted with only PC6 as a significant predictor and with r-squared of 0.07).

Note that the 1st principal component predicts significantly all ratings. Also, the effect size of this component is approximately the same for FLUENCY, PAUSING and SPEECH RATE assessments.

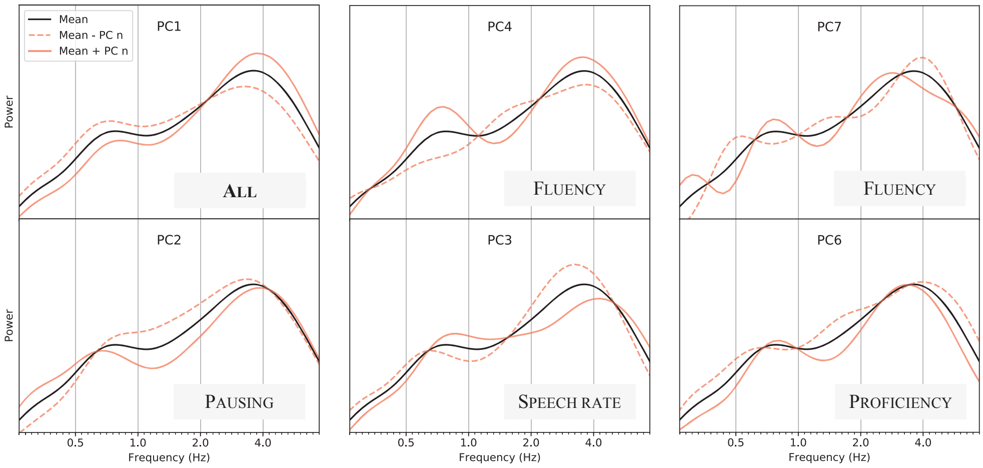
The shapes of the global spectra correspond best with SPEECH RATE assessments, explaining almost 40 % of variance, while the models for FLUENCY and PAUSING explain about 27 and 25 % of rating variance, respectively.

Interestingly, with the exception of the first principal component (PC1), different principal components (different aspects of global spectrum shapes) correspond to different L2 assessment types: PC4 and PC7 to FLUENCY, PC2 to PAUSING, and PC3 to SPEECH RATE.

Fig. 4 illustrates the effects of individual components on spectral shapes. PC1 shows that the ratings in all assessment categories improve for the spectra with lower ≈ 0.6 Hz and the peak at ≈ 4 Hz more prominent and shifted towards higher frequencies (the solid red line in the top left pane in Fig. 1). PC4 indicates that better FLUENCY ratings were given to recordings with spectra with both peaks higher and a more prominent gap between them, while PC7 associates better ratings with, again, more prominent ≈ 0.6 Hz peak but the ≈ 4 Hz peak shifted towards lower frequencies.

According to PC2 (bottom left pane in Fig. 1), PAUSING rating also improves with deeper gap be-

Figure 4: Effects of significant components. Black curves depict global spectrum averaged across all speakers. The solid red lines show effects of the components in the "positive" direction, i.e., the direction predicted by the models to increase the rating in the given assessment category; the dashed red lines show opposite "negative" effects. Assessment categories significantly explained by the given effects are marked for convenience.



tween the peaks and shifting the peaks further apart. PC3 shows that higher SPEECH RATE assessments were given for the renditions with the ≈ 4 Hz peak shifted towards the higher frequencies.

4. DISCUSSION

The presented wavelet analysis captures regularities present in the speech signal, namely the regularity associated with syllable production and one presumably linked to phrasing. These regularities are visible as two distinct peaks in wavelet power spectra at appropriate frequencies.

The peaks are clearly present for native, fluent speech for both analyzed languages. Interestingly, while syllable-related regularity remains discernible in spectra obtained from L2 speech signal, the peak associated with phrases is less pronounced. As seen in Fig. 1 this likely results from the irregularities in prosodic chunking in L2 speech where disfluencies, hesitations, etc., blur the rhythmic hierarchy. As our analysis shows, the position and prominence of these spectral peaks explains to a measurable degree the level of fluency as well as speech rate and pausing. The position of the "syllable rate" frequency component correlates with all assessment types: the higher the frequency, the higher the rating for fluency, pausing and speech rate. At the same time more fine-grained variation in the spectra relates to these three aspects separately. In this way, the method captures both the holistic overall perceptions of fluency as well as individual aspects of pausing and speech rate that participate and co-create this holistic perception. The fact that the spectral shape do not correspond with the proficiency ratings equally well might be due to different demands on proficiency ratings; one can be a reasonably fluent speaker of a foreign language without necessarily being very proficient in terms of, e.g., correct stress placement, pronunciation, etc. On the other hand it should be noted, that individual speech style or reading skills in L1 can affect L2 fluency [4].

The presented method is applied straight to the speech signal and requires no manual labelling of the signal. As such, it could be readily applied to assist in automatic fluency assessment. The viability of the approach would, however, need to be tested on speakers with different language backgrounds and greater and more widely spread L2 skill levels. Naturally, the applicability to less structured, spontaneous speech should also be evaluated.

Further evaluation of the method will entail comparing the most prominent frequency ranges with syllable, foot, and phrase durations manually extracted from the test data. For example, the higher frequency of the ≈ 4 Hz for Slovak compared to English data (see Fig. 2) might be related to syllable being the major rhythmic unit in Slovak and foot in English. This observation, if validated, might serve as a basis for a new method for comparing rhythmic characteristics of languages.

5. ACKNOWLEDGMENTS

We would like to thank Peter Kleman and Jana Beňušová for data collecting and assessments.

6. REFERENCES

[1] Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., De Jong, N. H. 2013. What makes speech sound fluent? the contributions of pauses, speed and repairs. *Language Testing* 30(2), 159–175.

[2] Cucchiarini, C., Strik, H., Boves, L. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *the Journal of the Acoustical Society of America* 111(6), 2862–2873.

[3] De Jong, N. H. 2016. Fluency in second language assessment. *Handbook of second language assessment* 203–218.

[4] De Jong, N. H., Groenhout, R., Schoonen, R., Hulstijn, J. H. 2015. Second language fluency: Speaking style or proficiency? correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics* 36(2), 223–243.

[5] Derwing, T. M., Rossiter, M. J., Munro, M. J., Thomson, R. I. 2004. Second language fluency: Judgments on different tasks. *Language learning* 54(4), 655–679.

[6] Eriksson, A., Suni, A., Vainio, M., Šimko, J. 2018. The acoustic basis of lexical stress perception. *Proc. 9th International Conference on Speech Prosody 2018* 70–74.

[7] Fulcher, G. 1996. Does thick description lead to smart tests? a data-based approach to rating scale construction. *Language Testing* 13(2), 208–238.

[8] Guion, S. G., Flege, J. E., Liu, S. H., Yeni-Komshian, G. H. 2000. Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics* 21(2), 205–228.

[9] Hahn, L. D. 2004. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly* 38(2), 201–223.

[10] Houtgast, T., Steeneken, H. J. 1985. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America* 77(3), 1069–1077.

[11] Kallio, H., Šimko, J., Huhta, A., Karhila, R., Vainio, M., Lindroos, E., Hildén, R., Kurimo, M. 2017. Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level. *AFinLA-e: Soveltavan kielitieteen tutkimuksia* (10), 193–213.

[12] Kallio, H., Suni, A., Virkkunen, P., Šimko, J. 2018. Prominence-based evaluation of l2 prosody. *Proc. Interspeech 2018* 1838–1842.

[13] Koponen, M., Riggenbach, H. 2000. Overview: Varying perspectives on fluency. *Perspectives on fluency*. University of Michigan 5–24.

[14] Kormos, J., Dénes, M. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32(2), 145–164.

[15] Lennon, P. 1990. Investigating fluency in efl: A quantitative approach. *Language learning* 40(3), 387–417.

[16] Leong, V., Stone, M. A., Turner, R. E., Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America* 136(1), 366–381.

[17] Loukina, A., Kochanski, G., Rosner, B., Keane, E., Shih, C. 2011. Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America* 129(5), 3258–3270.

[18] Munro, M. J., Derwing, T. M. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning* 45(1), 73–97.

[19] Skehan, P. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics* 30(4), 510–532.

[20] Suni, A., Šimko, J., Aalto, D., Vainio, M. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language* 45, 123–136.

[21] Suni, A. S., Šimko, J., Vainio, M. T., others, 2016. Boundary detection using continuous wavelet analysis. *Proceedings of Speech prosody 2016*.

[22] Torrence, C., Compo, G. P. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society* 79(1), 61–78.

[23] Trofimovich, P., Baker, W. 2006. Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech. *Studies in second language acquisition* 28(1), 1–30.

[24] Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., Lorenzi, C. 2017. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America* 142(4), 1976–1989.

[25] de Wet, F., Van der Walt, C., Niesler, T. 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51(10), 864–874.

[26] Zechner, K., Higgins, D., Xi, X., Williamson, D. M. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication* 51(10), 883–895.