Duration and Declination in L2 Reading

First Author[a], Second Author[b]
[a] *First Affiliation, Country*
[b] *Second Affiliation, Country*

Keywords — duration, declination, prosody, second language learning, computer assisted language assessment, CALA, speech modulation theory, global intonation contour, Support Vector Machine, SVM

1. Background

In the extensive computer assisted language learning (CALL) literature on L2 intonation, much attention has been paid to the teaching and learning of stress and pitch accents, terminal contours and rhythm, from both phonological [1] and phonetic [2] perspectives. The present study takes a complementary approach and focuses on computer assisted language assessment (CALA), proposing partial prosody support for testing, in teaching or self-study, through supervised machine learning (ML) of differences in L1 and L2 utterance duration and fundamental frequency (F0) slope ('declination', 'inclination' etc.), with a 'no difference' null hypothesis, rather than an approach based on the declination:duration ratio [3, 4], on pitch accents [5], or on global height and range parameters [6].

A CALA pipeline was constructed, from F0 estimation and modelling through descriptive statistics to ML classification of L2 and L1 productions, with the medium-term aim of providing assistance for grading L2 speakers with probabilistic information [7] on L1-likeness, using (among other techniques) Support Vector Machine (SVM) classification. In [8] a similar SVM method was applied to classifying personalities using F0 contours.

2. Data and Method

The L1 British English data are sourced from the IVIE corpus [9], and consist of 64 readings of 5 disjunctive questions ('OR-questions', chosen here for their complex 2-level intonation structure) by 13 female native speakers. The L2 data consist of 120 recordings of the same sentences by 24 female advanced Chinese EFL students with 12 years of school and university English, who are native speakers of Mandarin and Shaanxi dialect. The speakers are of the same sex so that the same F0 estimation parameters could be used for all participants.

The questions are: Q1, *Are you growing limes or lemons?*; Q2, *Is his name Miller or Mailer?*; Q3, *Did you say mellow or yellow?*; Q4, *Do you live in Ealing or Reading?*; Q5, *Did he say lino or lilo?* The sentences were recorded independently by the L2 students on equipment of their choice, mainly Praat [10] on laptops, a familiar testing scenario for them. Sampling rates varied, so recordings were resampled to 16 kHz. Initial and final silences were cropped to about 100ms and recordings were normalised to unit amplitude. Q4, which is voiced throughout, is used for illustration in this contribution. Several conspicuous L1-L2 differences are already apparent in Figure 1 (e.g. amplitude pattern, in addition to longer duration and flatter or uptrend global slope for L2).

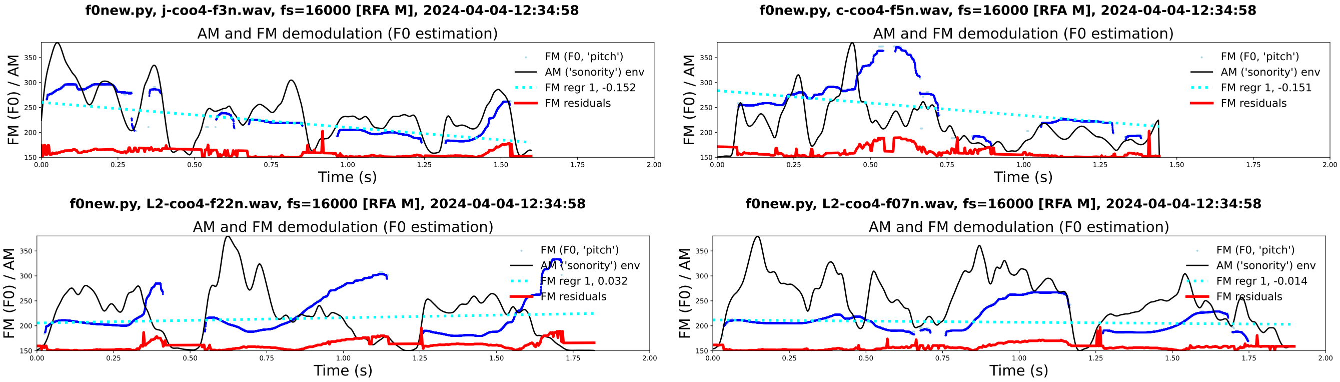

Figure 1: Q4, L1 (top), L2 (bottom), different readers: AM envelope, FM envelope, FM regression, residuals.

The methodological foundation is speech modulation theory [11], in which a carrier frequency is modulated by information signals, and which provides an integrated framework for all areas of phonetics. Modulation theory is as old as radio, and the terminology is the same as the labels on radio sets, but with a prosodic phonetic interpretation: FM (for frequency modulation of fundamental frequency, F0, in the larynx; a tone, pitch accent and intonation correlate) and AM (for amplitude modulation of speech sound and rhythm formants, by naso-oral filter; a sonority correlate):

$$\boldsymbol{Speech} = \boldsymbol{A_{AM}} A cos(2\pi(f + \boldsymbol{A_{FM}})t + \phi)$$

For example, rhythms use LF (low frequency) AM information signals, and intonation, tone and pitch accent use LF FM information signals, both with modulation frequencies which are in general below about 5 Hz. The focus in this study is on the FM information signal (f=frequency, t=time, A_{FM}=FM amplitude, A_{AM}=AM amplitude, ϕ=phase); phase is not treated further. The demodulation and modelling pipeline has the following steps:

1. AM demodulation: Amplitude envelope, extracted and smoothed, for visual time-frequency alignment.
2. Preliminary noise reduction: centre and peak clipping (10%).
3. Tuning: 3rd order Butterworth bandpass filter 120...380 Hz.
4. FM demodulation: custom time-domain F0 estimation with contour smoothing.
5. Model: linear regression line with interpolation of voicing gaps.
6. Feature extraction: residuals, and duration, slope, intercept values.

Table 1: Averaged results for L1 and L2 F0 contours.

Var:	Dur(s)	Slope	Intercept	SDF0	SDabsres
L1 mn:	1.807	-0.074	232.538	32.801	18.817
L2 mn:	2.629	-0.028	237.272	29.559	18.528
L1 SD:	0.348	0.051	26.040	9.988	4.623
L2 SD:	0.542	0.025	23.616	7.687	4.260
t-test:	$p < .01$	$p < .01$	$p < .01$	$p < .05$	$p > .05$
dur:x (corr):	-	0.507	-0.446	-0.484	-0.302

Selected global properties (duration, slope, intercept, SD of F0, SD of absolute values of residuals) were analysed for the L1 and L2 groups. Shapiro-Wilk tests showed near-normal distributions and T-tests were applied (Table 1). Duration, slope, intercept and SD F0 variables showed significant L1-L2 differences and thus refutation of the null hypothesis. The SD of absolute residuals did not. Slope correlated moderately with duration, confirming previous peakline studies [4]. Slope results also showed a tendency for steeper slopes in L1 than in L2 (see also Figure 1), a sociophonetic register factor. These differences found by basic statistical analysis suggest that a more general ML classification scheme could be used in order to discover whether the differences can be seen as a potential model for computer-assisted intonation grading, using a generalised target set rather than one individual, for example a teacher, as reference.

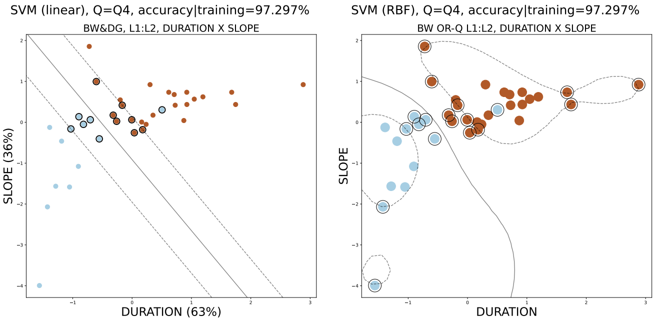

Figure 2: Q4 SVM graphs: linear (left), non-linear (right)

SVM classifiers were therefore trained for 2 classes and 2 features, with z-score standardised duration and slope as the features (Figure 2). The scatter plot already shows L1-L2 differences in duration and slope: skewed, variable downtrend slope, shorter duration for L1, but flat or uptrend slope, skewed, variable longer duration for L2. Duration contributes 63%, slope 36% (by coefficient weight) in Q4. SVM hyperplane accuracy is well above chance (full dataset: 90.22%), again refuting the 'no difference' null hypothesis: Q1: 94.44%, Q2: 81.08%, Q3: 83.78%, Q4: 97.30%, Q5: 91.89%. Both slope and duration thus differ between L1 and L2 in this dataset.

3. Discussion and conclusions

Duration and global F0 slope in L1 and L2 readings of disjunctive 'OR-questions' are distinguished with good accuracy in this small database, statistically and by SVM. Reasons for the difference, such as 20-year L1-L2 time lapse, L1 interference, L2 uncertainty, need further research. Practical applications will require larger and more varied datasets and more features [6], and large language models will become more relevant. However, distances of data vectors from the SVM hyperplane suggest one prospective component for a CALA assistant, showing tentative degrees of intonation proficiency as 'likelihood of prosodic native-likeness', with a generalised target set rather than a target individual.

References

[1] I. Mennen, "Beyond segments: Towards a l2 intonation learning theory," in *Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations*, E. Delais-Roussarie and M. A. . S. Herment, Eds. Springer Verlag, 2015, pp. 171–188.

[2] 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.

[3] D. Kocharov, N. Volskaya, and P. Skrelin, "F0 declination in Russian revisited," in *18th International Congress of Phonetic Sciences*, 2015.

[4] J. Yuan and M. Liberman, "F_0 declination in English and Mandarin broadcast news speech," *Speech Communication*, vol. 65, pp. 67–74, 2014.

[5] A. Rosenberg, "Classification of prosodic events using quantized contour modeling," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2020, p. 721–724.

[6] H. Ding, R. Hoffmann, and D. Hirst, "Prosodic transfer: a comparison study of F0 patterns in L2 English by Chinese speakers," in *Speech Prosody 2016*, 2016, pp. 756–760.

[7] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[8] U. D. Reichel, "Personality prediction based on intonation stylization," in *18th International Congress of Phonetic Sciences*, 2015.

[9] E. Grabe and B. Post, "Intonational variation in the British Isles," in *Proceedings of the International Conference on Speech Prosody*, P. Gilles and J. Peters, Eds., 2002.

[10] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[11] H. Traunmüller, "Conventional, biological, and environmental factors in speech communication: a modulation theory," *Phonetica*, vol. 51, no. 1-3, pp. 170–183, 1994.