

IDENTIFICATION OF NON-NATIVE ENGLISH SPEAKERS' L1s VIA PATTERNS OF PROSODIC FEATURE DEVIANCE FROM NATIVE SPEAKER NORMS

Jim Talley

Linguistic Computing Systems, Austin, TX USA
talley@lingcosms.com

ABSTRACT

This study is an exploratory data science-based look at the question of whether the first language (L1) of non-native speakers of English can be identified from only a few simple syllable- and utterance-level prosodic features of their speech. Simple machine learning (ML) modeling on these loudness, pitch, and duration cues yields imperfect, but much better than chance, discrimination (1) between each individual L1 and General American English (GAE), and (2) between the five studied L1s.

The described modeling is based upon “atypicality scores” (*a-Scores*) for the prosodic features, representing the degree to which features deviate, or not, from GAE native speaker norms. The prosodic features, their normalizations, and the *a-Score* characterizations are discussed.

Finally, ML-based feature selection analysis examines the individual prosodic features' relative importances for the individual L1 vs. GAE discrimination tasks and for the 5-way, forced-choice L1 classification task.

Keywords: prosody, L2 English, feature atypicality, machine learning, corpus phonetics

1 INTRODUCTION

Prominences are fundamental in speech, serving to mark focused words, to aid in speech segmentation, to support lexical access, to bootstrap first language (L1) learning [1], and more. There are multiple means employed by languages to signal prominence – increases in duration, significant changes in pitch, increases in loudness, and (non-)reduction of vowels are all markers of stressed syllables in English [2]. For an L2 learner of English to achieve native-like prosody, he/she must master all four of those aspects in dynamic combination (even in trading relations with each other [3]). English is, of course, not alone in richly marking prominent words/syllables, but in some languages, the four recognized correlates of stress that English uses may be “occupied” by other linguistic functions – for example, in languages with

lexical tone (*e.g.*, Vietnamese), speakers may be constrained with respect to pitch variations for stress marking, or in languages with phonemic length (*e.g.*, Finnish), duration changes to mark stress may lead to lexical confusions. This raises the question of whether ingrained prosodic habits related to the L1 of a non-native speaker (NNS) significantly, and predictably, influence his/her L2 prosody.

Such cross linguistic influence (CLI) certainly plays some role in L2 spoken language acquisition, though there has been robust discussion regarding the centrality or significance of that role [*e.g.*, 2,4,5]. This exploratory study is not an attempt to advocate for, or against, a strong position regarding the role of CLI. It merely asks, via a corpus and machine learning approach, if there are, indeed, reliably detectable manifestations of non-English L1 prosody on English L2 speech prosody.

2 THE DATA

The on-line Speech Accent Archive (SAA) [6] of English language recordings of the short “Stella” passage, read by thousands of native and non-native speakers, is a rich resource for addressing that question. This study selects SAA speakers from six L1s: General American English (**en**), Andean Latin American Spanish (**es**), Japanese (**ja**), Polish (**pl**), Russian (**ru**), and Mainland Mandarin (**zh**).

2.1 Speaker Selection

The SAA has an impressive variety of speakers, but, as a crowd-sourced resource, it also has considerable diversity with respect to the quality (and the equipment and environments) of recording. The L2 speakers' varying degrees of command of English is both an advantage and a challenge – false starts, self corrections, long pauses, and repetitions are abundant, which causes difficulties for automatic processing. Also, in SAA, L1s are not broken down regionally, so for example, native English speakers from the U.S., Ireland, Australia, India, England, etc. are lumped into a single English L1 pool, though metadata can be accessed to help tease them apart.

Most L1s in SAA have only a small set of speakers, and some of the more abundant languages suffer from the lack of homogeneity mentioned above. For this study, an average of 32 speakers were selected from the six L1s (**en, es, ja, pl, ru, & zh**)¹. Regional homogeneity was enforced, but there was not an extensive effort to balance genders. Recordings were further pre-screened for noise, echo, or generally poor quality, and speakers who spent their early school years in an English-speaking country were also filtered out. Prosodically typologically diverse languages – *e.g.*, lexical tone, syllable timing, phonemic length – was also of interest, in the sense that we might expect more differentiated CLI.

2.2 Features

Given a backdrop of interest in ESL/EFL pedagogical issues, a primary criterion in selection of prosodic features for this study was intuitiveness – that is, that they be amenable to actionable interpretation in a pedagogical context. There are a number of reasonably well-established prosodic metrics, *e.g.*, the openSMILE suite [7] and the PVI variants [8]. However, these were generally formulated for technical purposes, not for interpretability by/for general L2 learners. We have, therefore, chosen instead to use a small, relatively simple set of utterance and syllable characterizations (4 and 10, respectively) which can be explained to an L2 student (*e.g.*, IBSNGapDur is linking).

2.2.1 Basic features

The fourteen basic loudness (red), pitch (green), and duration (blue) features are listed and described in Table 1. Note that these (automatically extracted) basic features are normalized in several ways to compensate for cross-speaker (and recording) variability. First of all, units (mels and dB) are chosen to provide approximately linear scaling perceptually. Then, they are normalized against utterance means (and, in the case of pitch, also by standard deviation). These transformations are important for limiting idiosyncratic variance for modeling. The utterance level features, in turn, are intended to capture potentially significant aspects of between-speaker variation, such as pitch dynamism.

2.2.2 Atypicality scores (*a-Scores*)

In addition to the normalizations applied in deriving the basic prosodic features, we convert those basic

features to “atypicality scores” (*a-Scores*) for modeling by norming them against our population of GAE native speakers (NS). The *a-Scores* represent the degree to which prosodic features deviate (or not) from the normal variation observed among NS.















Sym	Short Name	Type	Description
	dBStdDev	Utt, Loud	Utterance loudness variability; more specifically, the stdev of the cNDur values from the utterance
	f0StdDev	Utt, Pitch	Utterance pitch variability; more specifically, the stdev of the utterance's syllables' nuclear f0 means (in mels)
	syllsPerSec	Utt, Dur	Speaking rate – the total utterance length (not including initial and final silences) divided by the number of syllables in the utterance
	pau2Spch	Utt, Dur	Utterance fluency – the ratio of silence time to speech time within the utterance (<i>i.e.</i> , not including initial and final silences)
	lDeltNLoud	Syll, Loud	Loudness change from the center syllable to its left neighbor, <i>i.e.</i> , $cNLoudDB_{i-1} - cNLoudDB_i$
	cNLoudDB	Syll, Loud	Mean loudness (in dB) computed over the nucleus of the i^{th} syllable, then divided by the mean of utterance's syllables' loudness values
	rDeltNLoud	Syll, Loud	Loudness change from the center syllable to its right neighbor, <i>i.e.</i> , $cNLoudDB_{i+1} - cNLoudDB_i$
	lDeltZNF0	Syll, Pitch	Pitch change from the center syllable to its left neighbor, <i>i.e.</i> , $cZNormF0_{i-1} - cZNormF0_i$
	cZNormF0	Syll, Pitch	Mean f0 (in mels) computed over the nucleus of the i^{th} syllable, then z-scaled using the utterance's f0 mean and stdev
	rDeltZNF0	Syll, Pitch	Pitch change from the center syllable to its right neighbor, <i>i.e.</i> , $cZNormF0_{i+1} - cZNormF0_i$
	IBSN-GapDur	Syll, Dur	Duration of the leftward between-syllable gap (<i>i.e.</i> , the length of pause, if any, between the $(i-1)^{th}$ and i^{th} syllables) divided by the mean duration of the utterance's syllables
	lDeltNDur	Syll, Dur	Duration change from the center syllable to its left neighbor, <i>i.e.</i> , $cNDur_{i-1} - cNDur_i$
	cNDur	Syll, Dur	Duration of the i^{th} syllable divided by the mean duration of the utterance's syllables
	rDeltNDur	Syll, Dur	Duration change from the center syllable to its right neighbor, <i>i.e.</i> , $cNDur_{i+1} - cNDur_i$

Table 1: Prosodic features used in this study

For v_x , a basic feature value from language x , we first calculate its normed value z_x (Equation 1) using (Yeo-Johnson transformed [9]) **en** sample statistics:

$$z_x = \frac{v_x - \mu_{en}}{\sigma_{en}} \quad (1)$$

Then, z_x is “soft capped” (logarithmically squashed) if its magnitude exceeds a threshold of $c=3$ standard deviations (Equation 2) to produce the *a-Score*, a_x :

$$a_x = \begin{cases} z_x & \text{if } -c \leq z_x \leq c \\ \text{sign}(z_x) \times (c + \log_e(1 + (|z_x| - c))) & \text{otherwise} \end{cases} \quad (2)$$

The logic behind the squashing is that more than $\pm c$ stddevs qualifies as very atypical, and, while we would still like monotonic increase (decrease) in value, any (possibly large and possibly spurious) amount beyond c should have limited influence in combinations with other features’ *a-Scores*.

3 MODELING

Exploration with respect to L1 transfer of prosodic characteristics is carried out via ML model training on the *a-Score* features. Ensembles of small, simple models (5 utterance and 77 syllable models, for 82 in all) are trained to produce independent likelihood estimates. The likelihoods are then fused (by simply averaging them) and converted into L1 classifications for the speakers (forced-choice – highest mean likelihood wins). Simple ML methods with few trainable parameters are used to avoid overfitting on the rather small data set.

All results mentioned in this paper are derived from linear regression estimators (LREs)² as implemented in Scikit-Learn [10]. The LRE models, importantly, are trained with sample weighting (to offset class size imbalances). Furthermore, in all cases, estimation is done using the leaving-one-out method – that is, when training/evaluating on a 200-item sample set, 200 models are trained, each being trained on 199 samples and each producing a single estimation for the one sample which was left out of its training set.

4 CLASSIFICATION RESULTS

We directly consider how distinguishable the 5 NNS L1s are from each other (given the *a-Score* featurization of the prosodic cues) via a 5-way forced-choice classification between them. The results (displayed in the normalized³ confusion matrix (CM) of Figure 1) are far from perfect at

58.04% correct overall, but also far above chance (~20%). Results are fairly uneven across the five L1s – **ru** is commonly misidentified as **pl** (and also **es**), and **ja** is confused as **zh** about a quarter of the time.

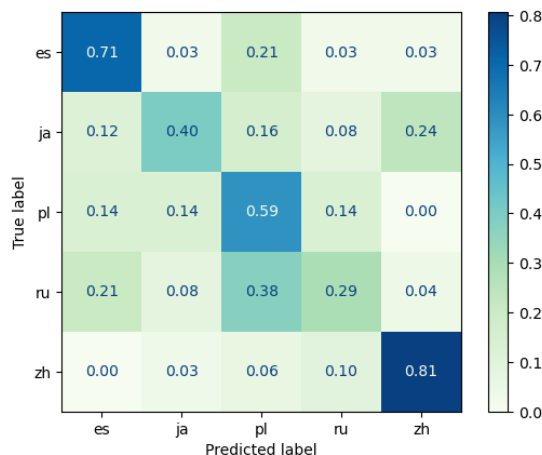


Figure 1: 5-way, forced-choice L1 classification

To test the conjecture that the “non-nativisms” of the L1s were interfering with each other, thus impacting achievable accuracy, we also modeled each L1 separately vs. **en** – *i.e.*, asking whether speakers’ L1s were reliably identifiable, vis-a-vis the native speakers, based on their prosody.

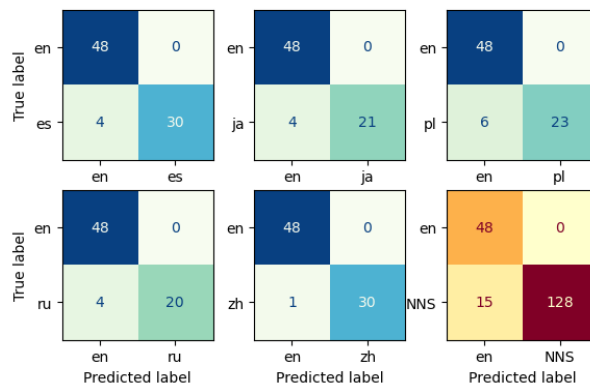


Figure 2: Separately distinguishing each NNS L1 from **en** (NS), plus **en** vs. the pooled NNS L1s

Figure 2 shows the results of the set of 5 binary L1 models (unnormalized CMs), plus the results of a binary model built from all of the NNSs pooled together vs. **en**. The **en**-NNS model was correct 92.15% of the time, while the **en**-L1 models averaged 95%.⁴ So, clearly, the prosodic features do support distinguishing the L1s from GAE (**en**).

5 FEATURE IMPORTANCE RESULTS

Given demonstration of the capacity to more or less identify L1s from these data, we would also like to

know what prosodic features actually matter for performing that task. One approach to estimating the relevance of the various prosodic cues (as manifested by the data features used in the study) is to look at how their presence, or absence, affects the ability to carry out an identification task. Ideally, we would like to consider every one of the possible combinations of features in that assessment⁵, but instead opted for the common, but sub-optimal, compromise of doing greedy search.

Greedy feature selection starts with single features – training models with them individually and picking the feature which provides the best performance on the task. It then extends to two features by considering remaining features in combination with the 1st ranked feature, and so on.

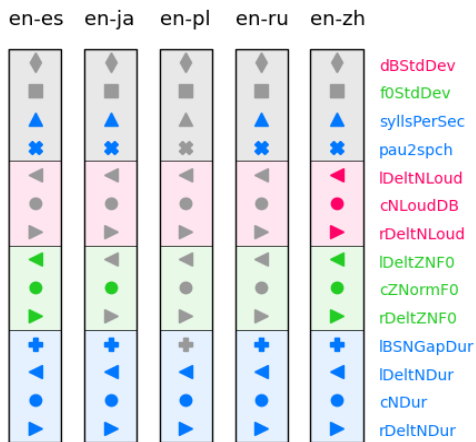


Figure 3: Features, by L1, which individually can discriminate the L1 from **en** at 80% or higher

It would be of definite interest to be able to derive robust, ordered lists of the most important features distinguishing each of the L1s from **en**. Unfortunately, there was too much of a ceiling effect with that analysis – they almost all approached 100% accuracy with only one or two features (with many features tied for the top rankings). Figure 3 is a compromise. It shows the features (colored) which are (individually) most effective for the **en**-L1 binary classifications. Note the predominance of (blue) duration features. Figure 3 also provides some potential insight with respect to performance on the 5-way classification task – *i.e.*, it suggests that there may be clashes among the various L1s' most informative features. Best performing **zh** notably has a number of unique (clash-free) top features.

Table 2 is an ordered list of the top features found for the 5-way L1 discrimination task. Once again, duration cues are prevalent, even though cross-L1 clashes may be limiting their effectiveness.

Rank	5-way, Forced-choice between L1s	
	Feature	%Corr
1	syllsPerSec	29.37
2	pau2Spch	37.76
3	f0StdDev	40.56
4	dBStdDev	36.36
5	cZNormF0	34.27
6	cNDur (rDeltNDur)	44.76
7	rDeltZNF0	52.44
8	rDeltNDur	55.94
9	lBSNGapDur	57.34

Table 2: Feature importance (greedy search)

It is probably unwise to put too much stock in the specific selection of features here. The automatically extracted data (from the challenging SAA recordings) are still fairly rough. To make more confident conclusions, we would want to work with a significantly larger, cleaner corpus.

6 SUMMARY

This study was primarily conducted to prototype a new methodology, including: 1) choosing a set of features with pedagogical potential; 2) defining a robust representation scheme for those features (the *a*-Scores); and, 3) validating an ensemble-of-limited-experts approach to small parameter space ML. The overarching question being asked of the corpus with this methodology was whether prosodic footprints of L1s are detectable in L2 English.

The task – roughly the equivalent of guessing the L1 of a speaker after hearing the melody of a subject's speech played on a piano – would likely be very challenging for humans. With noisy data, an impoverished representation of the speech (lacking, presumably useful, segmental information) and weak modeling, expectations for the study were low.

It is quite clear from the successful binary **en**-L1 and **en**-NNS results that non-native speakers are distinguishable from native speakers – *i.e.*, NNS footprints exist in the data. The 5-way forced-choice task speaks more directly to the question of whether characteristic L1 footprints also exist. That question is less conclusively answered. While the accuracy exceeded expectations, coming in well above chance, it was somewhat uneven across L1s and clearly more difficult than **en** vs. NNS classification.

7 REFERENCES

- [1] Gervain, J. 2022. Word frequency and prosody bootstrap basic word order in prelexical infants. *Speech Prosody* 2022.
- [2] Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. 2014. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, 57(4), 1468-1479.
- [3] Howell, P. 1993. Cue trading in the production and perception of vowel stress. *J. Acoust. Soc. Am.* (Oct) 94(4), 2063-2073.
- [4] Munro, M. 2018. How well can we predict second language learners' pronunciation difficulties? *The CATESOL Journal* 30.1, 267-281.
- [5] Wardhaugh, R. 1970. The Contrastive Analysis Hypothesis. *TESOL Quarterly*, 4, 123.
- [6] Weinberger, S. 2015. Speech accent archive. George Mason University. <http://accent.gmu.edu/>.
- [7] Coutinho, E., Hoenig, F., Zhang, Y., Hantke, S.; Batliner, A., Noth, E., & Schuller, B. 2016. Assessing the prosody of non-native speakers of English: Measures and feature sets. *LREC* 2016.
- [8] Low, E.L., Grabe, E., & Nolan, F. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43 (4), 377–401.
- [9] Yeo, I.K. & Johnson, R.A. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.
- [10] Pedregosa *et al.*, 2011. Scikit-learn: Machine Learning in Python, *JMLR* 12, 2825-2830.

¹ The selected SAA speaker numbers were: **en** (english<n>) – 9, 10, 32, 36, 39, 49, 51, 54, 55, 59, 60, 62, 63, 65, 68, 71, 74, 75, 78, 81, 82, 88, 90, 92, 96, 99, 106, 107, 114, 117, 118, 121, 123, 124, 126, 128, 133, 138, 143, 150, 157, 161, 162, 163, 165, 169, 171, 173; **es** (spanish<n>) – 3, 6, 7, 11, 13, 14, 18, 21, 23, 32, 42, 52, 53, 64, 67, 68, 70, 75, 78, 79, 80, 82, 83, 87, 88, 96, 97, 100, 102, 108, 111, 115, 123, 124; **ja** (japanese<n>) – 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29; **pl** (polish<n>) – 1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 31, 32, 34, 35, 36; **ru** (russian<n>) – 1, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 24, 30, 34, 36, 37, 41, 45, 47, 49, 50; **zh** (mandarin<n>) – 1, 2, 3, 4, 5, 6, 8, 14, 17, 19, 21, 22, 25, 26, 28, 31, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 51, 52, 53

² Experiments were done with other modeling methods (k-Nearest Neighbors regressors, small neural networks (MLPs),...) but none were found to give accuracy advantages over the linear models, and training/evaluation time for the linear models was the fastest of the considered ML methods.

³ The confusion matrix of Table 1 is normalized to fractions of the true speakers (rows).

⁴ Individual **en**-L1 accuracies: **en-es**=95.12%, **en-ja**=94.52%, **en-pl**=92.21%, **en-ru**=94.44%, and **en-zh**=98.73%

⁵ Given the structure of the modeling in this work, the 5-way, forced-choice task alone would require training ~1.4B models for exhaustive feature combination exploration, while greedy search reduces that to a more manageable ~7.3M models trained.