An automated pipeline for preprocessing spontaneous L2 English prosody

Sylvain Coulange[1,2,3], Tsuneo Kato[3], Solange Rossato[2], Monica Masperi[1]

[1]*Univ. Grenoble Alpes, Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), 38000 Grenoble, France*
[2]*Univ. Grenoble Alpes, CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), 38000 Grenoble, France*
[3]*Doshisha Univ., Spoken Language Processing Laboratory (SLPL), 610-0394 Kyoto, Japan*
sylvain.coulange@univ-grenoble-alpes.fr, tsukato@mail.doshisha.ac.jp,
solange.rossato@univ-grenoble-alpes.fr, monica.masperi@univ-grenoble-alpes.fr

While numerous tools address L2 pronunciation, they tend to focus on segmental deviations, often neglecting prosody and lacking pedagogical feedback (Coulange 2023). In contemporary L2 speaking classes, the foremost priority is achieving "understandability," encompassing both being understood and achieving it as effortlessly as possible (commonly referred to as intelligibility and comprehensibility, Derwing and Munro 2015). Within this framework, assessing speech requires identifying phenomena that significantly hinder listener understanding, and prioritizing them in assessment. Pinpointing these target areas in students' speech facilitates focused improvement for enhanced comprehensibility.

In the realm of English as a foreign language, rhythm, notably the placement of hesitation markers like silent or filled pauses and lexical stress realization, plays a crucial role in comprehensibility. Conversely, common segmental, grammatical, and lexical deviations show a comparatively lower impact on the cognitive load associated with speech processing, though they remain important considerations (Isaacs, Trofimovich, and Foote 2018; Walker, Low, and Setter 2021; Tortel 2021, among others). While some tools analyze pause frequency and length (de Jong, Pacilly, and Heeren 2021) or classify lexical stress (Ferrer et al. 2015; Shahin, Epps, and Ahmed 2016), our investigation identified a gap in tools considering the syntactic context of pauses and the degree of contrast between stressed and unstressed syllables. We developed a fully automated pipeline for processing spontaneous L2 English speech, that analyzes pausing and stress patterns.

Two releases of the Pauses and Lexical Stress Processing Pipeline[1] (plspp) currently coexist. Both are based on WhisperX speech recognition (Bain et al. 2023), but the first one (plspp1) extracts stress related acoustic parameters from syllable nuclei points, while the second version (plspp2) uses an extra layer of phoneme-level forced alignment using Montreal Forced Aligner (McAuliffe et al. 2017) and extracts acoustic parameters within vowel intervals. Pause pattern analysis is based on inter-word intervals' duration, part-of-speech context, opening and closing constituents, considering their size and syntactic depth (Kitaev, Cao, and Klein 2019). Pauses lower and upper duration thresholds can be easily set up to consider only intervals of a certain duration.

The analysis of lexical stress involves comparing word-level prosodic shapes with their expected stress pattern extracted from a reference dictionary, and measuring the prosodic contrast between stressed and unstressed syllables. Each syllable is represented by three speaker-normalized measures: F0, intensity and duration.

[1] The pipeline is open-source and freely available here https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp.

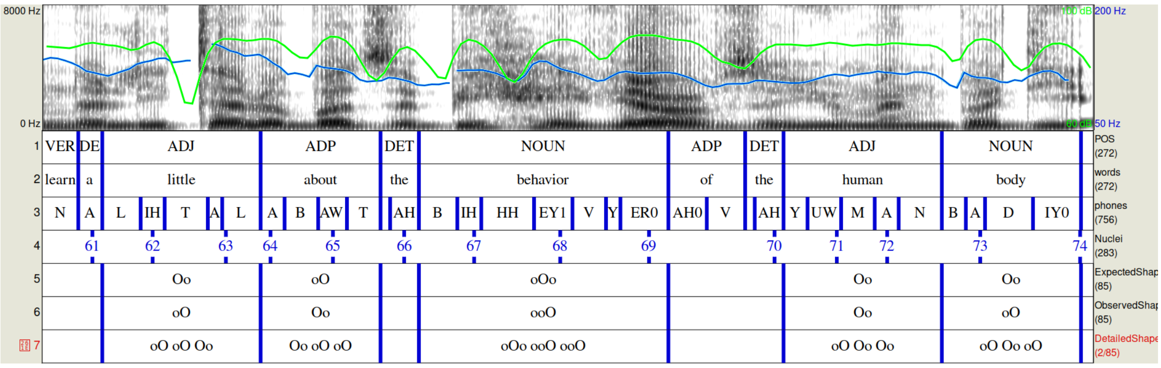

8000 Hz
0 Hz
100 dB
60 dB
200 Hz
50 Hz
1 VER DE ADJ ADP DET NOUN ADP DET ADJ NOUN — POS (272)
2 learn a little about the behavior of the human body — words (272)
3 N A L IH T A L A B AW T AH B IH HH EY1 V Y ER0 AH0 V AH Y UW M A N B A D IY0 — phones (756)
4 61 62 63 64 65 66 67 68 69 70 71 72 73 74 — Nuclei (283)
5 Oo oO oOo Oo Oo — ExpectedShap (85)
6 oO Oo ooO Oo oO — ObservedShap (85)
7 oO oO Oo Oo oO oO oOo ooO ooO oO Oo Oo oO Oo oO — DetailedShape (2/85)

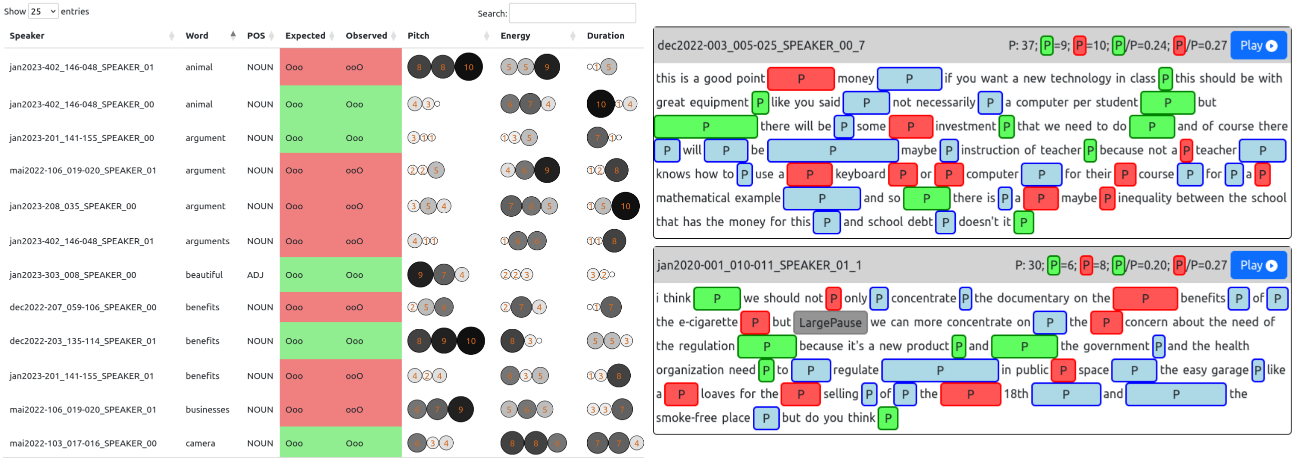
Figure 1: Example of a TextGrid output from plspp2 showing POS tags (1), transcribed text (2), phoneme alignment (3), syllable nuclei (4), expected prosodic shape (5), observed prosodic shape (6), F0, intensity and duration shapes (7)

Acoustic stress is inferred to be the most prominent syllable within the word for each dimension, and these three dimensions are merged with equal weight to obtain a single global representation easier to handle. Stress position is analyzed through a binary representation of syllables, with "O" representing the stressed syllable and "o" representing the other syllables in the word. Both releases do not consider the secondary stress yet.

Both versions output several tables including one listing the stressed words with their acoustic detailed information, and another table listing all inter-word intervals along with their duration and syntactic context for pause pattern analysis. Moreover, a TextGrid file is generated for each audio file allowing further acoustical analysis (cf. Figure 1).A visualisation tool is also being developed in order to more easily overview – and dive in – the results. This tool exists as a light stand-alone html/js-only version encompassed in the plspp pipeline; as well as a Django server-based application for web hosting purposes.

This pipeline has already been used in several studies involving French, Japanese and Korean learners of English, as well as native speakers of English, in elementary school and at university, on spontaneous, recited or read aloud speech.

Our presentation will describe how both pipeline work and elucidate the decision-making process behind them, thereby initiating a discussion about their inherent limitations and possible future improvements. Additionally, we will showcase the different ongoing studies, presenting preliminary results that have been obtained thus far.



Show 25 entries Search:

Speaker	Word	POS	Expected	Observed	Pitch	Energy	Duration
jan2023-402_146-048_SPEAKER_01	animal	NOUN	Ooo	ooO	8 8 10	5 5 9	○ 1 5
jan2023-402_146-048_SPEAKER_00	animal	NOUN	Ooo	Ooo	4 3 ○	6 7 4	10 1 4
jan2023-201_141-155_SPEAKER_00	argument	NOUN	Ooo	Ooo	3 1 1	1 3 5	7 1 ○
mai2022-106_019-020_SPEAKER_01	argument	NOUN	Ooo	ooO	2 2 5	4 6 9	1 2 8
jan2023-208_035_SPEAKER_00	argument	NOUN	Ooo	ooO	3 5 4	7 6 5	1 5 10
jan2023-402_146-048_SPEAKER_01	arguments	NOUN	Ooo	ooO	4 1 1	1 6 6	1 1 8
jan2023-303_008_SPEAKER_00	beautiful	ADJ	Ooo	Ooo	9 7 4	2 2 3	3 2 ○
dec2022-207_059-106_SPEAKER_00	benefits	NOUN	Ooo	ooO	2 5 6	2 7 4	○ 1 7
dec2022-203_135-114_SPEAKER_01	benefits	NOUN	Ooo	Ooo	8 9 10	8 3 ○	5 5 3
jan2023-201_141-155_SPEAKER_01	benefits	NOUN	Ooo	ooO	4 2 4	6 3 5	1 3 8
mai2022-106_019-020_SPEAKER_01	businesses	NOUN	Ooo	ooO	6 7 9	5 6 5	3 3 7
mai2022-103_017-016_SPEAKER_00	camera	NOUN	Ooo	Ooo	6 3 4	8 8 6	7 7 4

dec2022-003_005-025_SPEAKER_00_7 P: 37; P=9; P=10; P/P=0.24; P/P=0.27 Play

this is a good point P money P if you want a new technology in class P this should be with great equipment P like you said P not necessarily P a computer per student P but P there will be P some P investment P that we need to do P and of course there P will P be P maybe P instruction of teacher P because not a P teacher P knows how to P use a P keyboard P or P computer P for their P course P for P a P mathematical example P and so P there is P a P maybe P inequality between the school that has the money for this P and school debt P doesn't it P

jan2020-001_010-011_SPEAKER_01_1 P: 30; P=6; P=8; P/P=0.20; P/P=0.27 Play

i think P we should not P only P concentrate P the documentary on the P benefits P of P the e-cigarette P but LargePause we can more concentrate on P the P concern about the need of the regulation P because it's a new product P and P the government P and the health organization need P to P regulate P in public P space P the easy garage P like a P loaves for the P selling P of P the P 18th P and P the smoke-free place P but do you think P

Figure 2: Overview of the stress pattern analysis (left) and pause patterns (right), with inter-clause pauses in green, inter-phrase in blue and intra-phrase in red

References.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman (2023). "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *Interspeech*.

Coulange, Sylvain (2023). "Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up". In: *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*. Ed. by Alice Henderson and Anastazija Kirkova-Naskova, pp. 11–22. DOI: 10.5281/zenodo.8137754.

de Jong, Nivja H., Jos Pacilly, and Willemijn Heeren (2021). "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically". In: *Assessment in Education: Principles, Policy & Practice* 28.4, pp. 456–476. DOI: 10.1080/0969594X.2021.1951162.

Derwing, Tracey M. and Murray J. Munro (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Ferrer, Luciana, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda (2015). "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems". In: *Speech Communication* 69, pp. 31–45. DOI: https://doi.org/10.1016/j.specom.2015.02.002.

Isaacs, Talia, Pavel Trofimovich, and Jennifer Ann Foote (2018). "Developing a user-oriented second language comprehensibility scale for English-medium universities". In: *Language Testing* 35.2, pp. 193–216. DOI: 10.1177/0265532217703433.

Kitaev, Nikita, Steven Cao, and Dan Klein (2019). "Multilingual Constituency Parsing with Self-Attention and Pre-Training". In: *ACL*. Florence, Italy, pp. 3499–3505.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Proc. Interspeech 2017*, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386.

Shahin, Mostafa Ali, Julien Epps, and Beena Ahmed (2016). "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning". In: *Interspeech*, pp. 175–179.

Tortel, Anne (2021). "Le rythme en anglais oral : considérations théoriques et illustrations sur corpus". In: *Recherche et pratiques pédagogiques en langues - Cahiers de l'APLIUT* Vol. 40 N°1. DOI: 10.4000/apliut.8857.

Walker, Robin, Ee-Ling Low, and Jane Setter (2021). *English pronunciation for a global world*. Oxford University Press.