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

Context

- Prosody (e.g. fluency and rhythm) is crucial for intelligibility in L2 [1].
- Little work is done on prosody in classrooms.
- Teachers and learners both need tools helping measuring prosody.
- Existing Computer Assisted Pronunciation Training systems rarely deal with spontaneous speech and conversation situations.

Proposition

- We developed a pipeline to help measuring speech segmentation and rhythm in spontaneous conversations involving learners of English.
- It combines most recent tools for speech recognition and syntactic analysis with lexical stress analysis and pause position analysis.

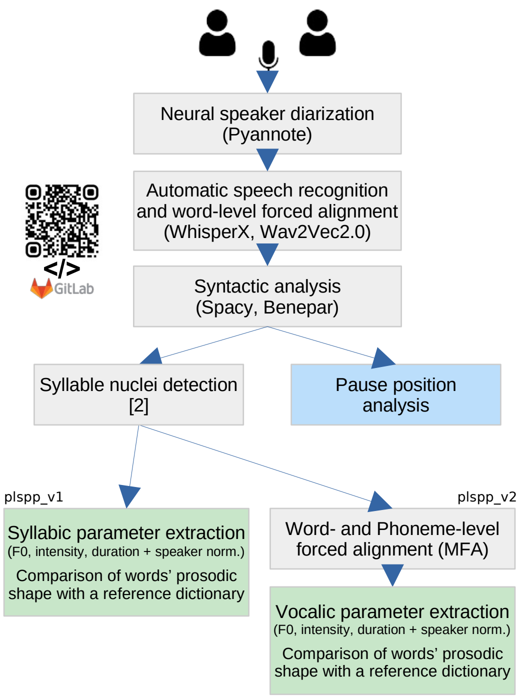
Input

- Multispeaker spontaneous L2 English audio recordings.

Outputs

- TextGrid and csv files showing syllabic prominence, expected stress patterns and pause syntactic context.
- Visualisation of stress and pauses through an interactive web interface.

The Pauses & Lexical Stress Processing Pipeline (PLSPP)



Neural speaker diarization (Pyannote)

Automatic speech recognition and word-level forced alignment (WhisperX, Wav2Vec2.0)

Syntactic analysis (Spacy, Benepar)

Syllable nuclei detection [2]

Pause position analysis

plspp_v1

Syllabic parameter extraction (F0, intensity, duration + speaker norm.)

Comparison of words' prosodic shape with a reference dictionary

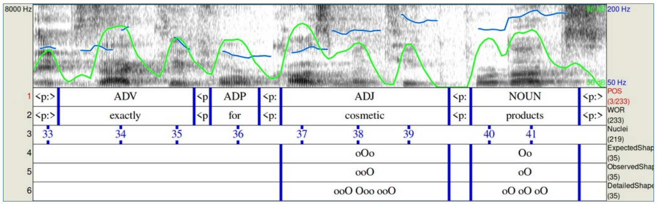
plspp_v2

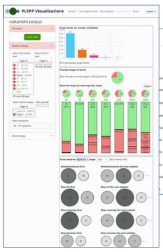
Word- and Phoneme-level forced alignment (MFA)

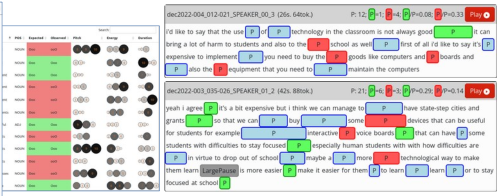
Vocalic parameter extraction (F0, intensity, duration + speaker norm.)

Comparison of words' prosodic shape with a reference dictionary

Insight of a TextGrid output (plspp_v1)







dec2022-004_012-021_SPEAKER_00_3 (26s. 64tok.) P: 12; P=1; P=4; P/P=0.08; P/P=0.33 Play

i'd like to say that the use P of P technology in the classroom is not always good P it can bring a lot of harm to students and also to the P school as well P first of all i'd like to say it's P expensive to implement P you need to buy the P goods like computers and P boards and P also the P equipment that you need to P maintain the computers

dec2022-003_035-026_SPEAKER_01_2 (42s. 88tok.) P: 21; P=6; P=3; P/P=0.29; P/P=0.14 Play

yeah i agree P it's a bit expensive but i think we can manage to P have state-step cities and grants P so that we can P buy P some P devices that can be useful for students for example P interactive P voice boards P that can have P some students with difficulties to stay focused P especially human students with with how difficulties are P in virtue to drop out of school P maybe a P more P technological way to make them learn LargePause is more easier P make it easier for them P to learn P learn P or to stay focused at school P

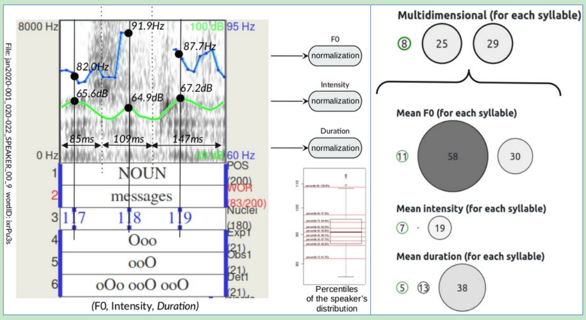
A server-based visualisation tool allows to easily view the processing outputs.

Pipeline Evaluation & Limitations

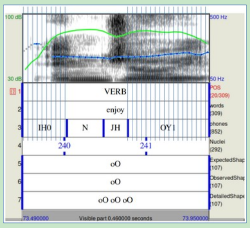
- Precision of ASR and syntactic analysis varies greatly with speaking tasks and speakers' proficiency level. Because of MFA sensitivity to disfluencies, plspp_v2 is less robust than v1 and works well only with read speech so far.
- With spontaneous conversations from [3] (plspp_v1):
  - 41% of words have adequat number of syllable nuclei detected and thus analysed.
  - Manual evaluation of random 100 target words showed that 17% were miss-recognized or miss-aligned.
- With read-aloud monologues from [4] (plspp_v2 with forced-alignment of reference texts):
  - 100% of words are analysed since it does not depend on syllable nuclei detection.
  - Manual evaluation of 400 words among 8 Japanese-L1 and 8 native speakers showed that 8.8% of words were miss-aligned (Japanese-L1: 9%, native speakers: 8.5%).

Lexical Stress Analysis

- plspp_v1: Lexical stress is estimated from prosodic measures on syllable nuclei, based on F0, intensity and syllable duration.
- plspp_v2: Lexical stress is estimated from prosodic measures of vowel intervals, based on mean F0, max intensity and vowel duration.



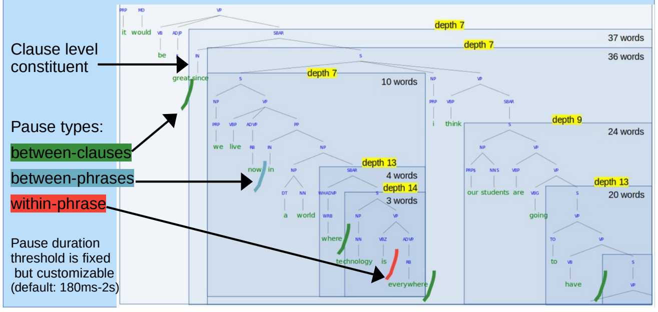
8000 Hz
91.9Hz
100 dB
95 Hz
87.7Hz
82.0Hz
65.6dB
64.9dB
67.2dB
85ms
109ms
147ms
0 Hz
60 Hz
NOUN
messages
Ooo
ooO
oOo ooO ooO
(F0, Intensity, Duration)
F0
normalization
Intensity
normalization
Duration
normalization
Percentiles of the speaker's distribution
Multidimensional (for each syllable)
Mean F0 (for each syllable)
Mean intensity (for each syllable)
Mean duration (for each syllable)
25
29
58
30
19
38



- In both versions, each prosodic measure is converted in speaker percentile, so that 50 means median prominence level for any speaker, 0 is minimum and 100 is maximum prominence level.

Pause Position Analysis

- Pauses are categorized depending on their syntactic position in relation to clause and phrase level constituents, along with POS context, syntactic depth and nb. of words of adjacent constituents.



Clause level constituent

Pause types:

between-clauses

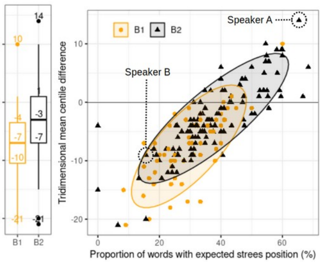
between-phrases

within-phrase

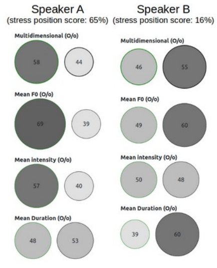
Pause duration threshold is fixed but customizable (default: 180ms-2s)

Stress accuracy and pause positions in B1/B2 French-L1 spontaneous conversations [3]

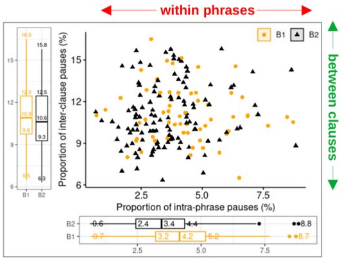
- Corpus: 176 French-L1 university students (11 hours), 6350 polysyllabic target words, 21942 pauses.
- Main observations:
  - Mean stress position accuracy varies greatly among speakers (0~68.4%, mean: 35.4%).
  - B2 speakers perform better than B1 in terms of stress position accuracy (36% vs. 29.6%, rank test p<.001) and prosodic contrast between expected primary stress syllable and mean of other syllables (p<.001).
  - Strong impact of last syllable lenghtening and pitch rise, the better the speaker mean stress position accuracy, the higher pitch and intensity of expected stressed syllable.
  - B2 speakers tend to make more but shorter pauses than B1 (median 34.3 pauses/min/speaker vs. 30.7, p<.01; 592ms vs. 615ms, p<.01), but proportionally less pauses within phrases than B1 speakers (4.2 pauses for 100 tokens vs. 3.4, p<.01).



Stress position accuracy (x) and prosodic contrast (y) per speaker



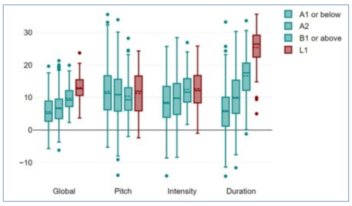
Prosodic contrast between expected primary stress (left) and other syllables (right)



Proportion of pauses per speaker (pauses/tokens ratio)

Prosodic contrast between content and function words in Japanese and English-L1 monologues [4]

- Corpus: 42 Japanese-L1 speakers (A1-B2), 9 English-L1 professional narrators, 34 hours read-aloud speech.
- PLSPP_v2, extended to monosyllabic words analysis, based on forced alignment of reference text.
- Main observations:
  - Contrast between most prominent syllable from content words and that of function words is higher for Native speakers than Japanese-L1 speakers (p<.001).
  - The higher the proficiency level, the bigger the contrast.
  - Most of this contrast is due to vowel duration parameter.



Mean prosodic contrast between content and function words

References:

[1] Levis, J. M. (2018). Cambridge applied linguistics: Intelligibility, oral communication, and the teaching of pronunciation.

[2] De Jong, N. H., Pacilly, J., Heeren, W. (2021) "Praat scripts to measure speed fluency and breakdown fluency in speech automatically." Assessment in Education: Principles, Policy & Practice, 28, 456-476.

[3] Coulange S, Kato T, Rossato S, Masperi M. (2024). Enhancing Language Learners' Comprehensibility through Automated Analysis of Pause Positions and Syllable Prominence. Languages 9(3):78

[4] Coulange, S., Nakanishi, M. (subm.). Measuring speech rhythm through automated analysis of syllabic prominences. Prosodic features of language learners' fluency (Speech Prosody WS), July 1, Leiden.

financé par IDEX Université Grenoble Alpes

科研費 KAKENHI This research is supported by JSPS Kakenhi n°23H00648