Investigating speaker pronunciation variability in speech embeddings: speaker and L1 effects on French as a Second Language

Maxime Fily, Martine Adda-Decker, Guillaume Wisniewski
INALCO, Université Sorbonne Nouvelle, Université Paris Cité
2 rue de Lille 75007 Paris, 4 rue des Irlandais 75005 Paris, 8 rue Albert Einstein 75013 Paris
maxime.fily@inalco.fr, martine.adda-decker@sorbonne-nouvelle.fr, guillaume.wisniewski@u-paris.fr

Abstract

Speech variation between native and non-native speakers of French is addressed with a low-resource method based on a frame-wise comparison of wav2vec2 acoustic embeddings, using fine-grained phonetic transcriptions by expert annotators as baseline. z-normalisation and t-normalisation are explored to assess what the embeddings contain in terms of phonetically analysable information. We explore non-supervised methods for solving basic speech-related research questions. Adapting Dynamic Time Warping to speech embeddings, we compare phonologically similar recordings of sentences read-aloud by native vs. non-native speakers of French. The question is whether XLSR-53 embeddings are more robust than MFCCs to inter-speaker vs. intra-speaker variability for different occurrences of the same words. Then we investigate whether native speaker productions are more or less stable than those of non-native speakers. Results suggest that the model allows phonetically meaningful correlative analyses. Working on the raw embeddings shows however that the representations are not speaker-independent, so with a view to address issues in relationship with L2 pronunciation variability, we show that t-normalisation brings us a way to separate fluency and accuracy effects in L2-speech. This shows that wav2vec2 encapsulates time-dependent phonetic information in the embeddings, including speaker accent which can not easily be disentangled from other speaker-specific characteristics.

Keywords: Cosine similarities, speech, L2-acquisition, Query-by-Example Spoken Term Detection, unsupervised method

1. Introduction

Despite the extensive instrumental and analytical resource available in phonetics, studies of pronunciation acquisition often still depend on impressionistic criteria, including native-likeness, accentedness, and intelligibility judgments (Saito et al., 2016). It is as if the immense variability inherent to foreign-accented speech[1] led researchers to resort only to perceptual criteria with native judges. It is argued in Machine-Learning ASR that a mere increase in the amount of training data "dilutes" speaker characteristics, thereby offering better robustness to variability, but recent findings (Zee et al., 2024) claim, on the contrary, that without attention to corpus balance, more training data aggravates biases because the proportion of data from dominant categories increases more than from *minorised* groups. Large Audio Models (LAM) therefore encapsulate more bias in the representations, especially when the data is scraped across the internet rather than collected with controlled metadata and explicit consent (some evidence in Névéol et al., 2022; Gebru, 2019; Buolamwini and Gebru, 2018). For these groups, mass scraping results in less performant ASR, so is the case for people with a foreign accent (Khandelwal et al., 2020).

[1] driven by social factors, region, personal history. (See Moyer, 2013).

This study examines whether phonetic information is encoded in one multilingual (`XLSR-53`) and one monolingual (`wav2vec2-FR-7K-large`) `wav2vec2` (Baevski et al., 2020) model by comparing distances between embeddings generated from occurrences of the same words. With this method, we then move on to measuring pronunciation variability on read-aloud speech through the lens of (i) inter-speaker vs. intra-speaker variability and (ii) L2-speech vs native speech. For this part, we focus the analyses on `XLSR-53` (Conneau et al., 2021) to measure the sensitivity of a multilingual model to speaker characteristics and native language.

The method uses a corpus of read sentences in native-French and L2-French. We start by assessing the correlation between the variability observed on `wav2vec2` (Baevski et al., 2020) neural model embeddings and the associated MFCCs (Mean-Frequency Cepstral Coefficients, a direct measurement on the audio signal). This paper also investigates the amount of speaker-specific information present within `wav2vec2` speech embeddings in comparison with the MFCCs. Based on a Query-by-Example approach using Dynamic Time Warping on speech embeddings, speaker-specific information/pronunciation is evaluated by measuring how robust the alignment is when speaker ID and foreign accents vary.

Our contribution consists in an unsupervised

Proceedings of SPEAKABLE @ LREC 2026, pages 86–97
11 May 2026.

method to measure pronunciation variation using `wav2vec2` embeddings. It is tested on `XLSR-53` multilingual model.

2. Existing methods

This section lays out the relevant literature on foreign accents in L2-speech, in terms of experimental methods and NLP frameworks.

2.1. Experimental methods in foreign accent evaluation

Pronunciation learning theories, despite their early development (e.g., Flege, 1981, 1995), have been challenging to prove empirically and therefore remain rarely tested in L2 acquisition frameworks (Kennedy and Trofimovich, 2017), even less so with specific, objective approaches. Pronunciation acquisition evaluation frameworks need non-judgmental feedback to help learners' derive a critical but constructive view of their own productions (Suzukida, 2021). Among recent reviews of L2-acquisition research, (Derwing, 2008) still deplore that at this point, many of the publications have not yet exploited the potential benefits of new technologies, in particular in "measuring progress", which is still mostly perception-based (although a few publications in experimental phonetics outline the role of direct measurements in L2-acquisition improvement, e.g., for retroflex consonants (Bliss et al., 2018) in English or for palatalized consonants in Russian (Lecocq, 2021)). In summary, a lot of the feedback to learners basically amounts to "did I do good?" without necessarily knowing what "doing good" means or which aspects of pronunciation need to improve. In a recent study, (Saito et al., 2016) devised a range of measurable variables to evaluate what participates to native-likeness in L2 speech. Their results identify speech rate among first order parameters for nativelikeness, but they focused quasi-exclusively on lexical variables (lemma, morphology, polysemy). They did not include phonetic accuracy of the output among the parameters although several studies show that when rating native-likeness in L2 speech, expert listeners scores correlate with the character error rate calculated between the narrow transcription of the segment and its native-like phonological transcription (Munro, 2008).

2.2. NLP methods in foreign accent evaluation

Large Language Models, which train on massive datasets, "exhibit and amplify stereotypical bias" (Ducel et al., 2025), as illustrated in Tatman (2017) where different genders and accents do not transcribe equally well using the (proprietary) YouTube ASR system. Efforts have been made to offer less biased ASR systems for the minorised languages (Havard et al., 2025), or to remove biases in Large Language Models via adversarial training. In low-resource scenario cases where resources are too scarce for such advanced debiasing methods, evaluating WER values in an ASR task can help identify the root causes for the biases before potentially offering solutions (Feng et al., 2021; Zhang et al., 2022).

Bartelds et al. (2022) estimated Foreign accent on non-native American English (Weinberger and Kunath, 2011) and concluded on a correlation between LAM embeddings and a nativelikeness assessment by humans. By comparing their results with how Levenshtein Distance (LD) or MFCCs correlate with nativelikeness judgment, they obtain results which are on average similar to LD or MFCCs. Their study is based on evaluating similarities between segments based on the Dynamic Time Warping (DTW) approach (Giorgino, 2009). It establishes a link between alignment performance and nativelikeness judgment, regardless of how the alignment is done: on LAM embeddings, MFCCs, or LD.

In this study, we are interested in avoiding nativelikeness judgments by instead correlating embedding alignments with two different metrics: WER and MFCC alignment cost. Query-by-Example Spoken Term Detection (QbE-STD), a long-standing word retrieval method which typically searches the minimal alignment cost between an audio sample and an audio lexicon and by doing so retrieves the correct dictionary entry. It can be applied to either audio signal (Le Ferrand et al., 2020) or speech embeddings (San et al., 2021). QbE-STD has been used in document retrieval on large audio archives (Ram et al., 2020; Hazen et al., 2009), or as an alternative to fine-tuning for the retrieval of low-resource language utterances when there is insufficient data (typically less than 2h. See Guillaume et al., 2022) for a proper fine-tuning (San et al., 2021).

Among recent studies using frame-wise, time-dependent embeddings, we cite the example of how QbE-STD has been done on low-resource languages such as Kunwinjku, an Australian Aboriginal language and Mboshi, a Bantu Congo Brazzaville language (Le Ferrand et al., 2020): the goal was to enable speakers of these oral languages to retrieve words from a dictionary. Among the acoustic features tested (including but not limited to MFCCs and fine-tuned `w2v2-en` embeddings, MFCCs gave out the best results, but were quite predictably very speaker-dependent. z-normalising the features (Lobanov, 1971; Xie and Jaeger, 2020) improved the recall values but the approach was only

applied to MFCC coefficients since z-normalising did not have any visible effect on the embeddings.

3. Research objective

Our study explores whether or not phonetic/pronunciation information is represented in two `wav2vec2` models (`XLSR-53` and `wav2vec2-FR-7K-large`) and if so, how they encode this information. We compare several occurrences of the same target words by measuring distances between them in the models' representations space. This approach consists in (i) verifying how well audio and embeddings are correlated after forced alignment at word level, and then, focusing on `XLSR-53` exclusively, (ii) assessing `XLSR-53` sensitivity to basic linguistic questions. We first verify whether different occurrences of a same speaker align more or less easily than those of different speakers. Then, the effect of speaker L1 (French or Russian) is assessed via cross comparisons of the alignment cost. Effects of embeddings z-normalisation[2] and t-normalisation are also assessed, always with a view to reducing the biases due to speaker ID or non-native speech in corpora.

By controlling the recording conditions for all the participants and by applying to speech embeddings an approach derived from experimental phonetics methods in an NLP framework, we propose to compare distance measurements in the neural representations space and in the audio space. We are interested in understanding to what extent it allows us to draw conclusions on basic/intuitive research questions: can we measure inter - vs. intra-speaker variability from the embeddings and is native speech more stable than L2 speech? These questions are already addressed widely in the literature but not yet with a frame-wise dual-space approach. Our contribution ultimately aims at establishing a correlation between distance measurements on two signals and their discrepancies in terms of Character Error Rate.

4. Material and method

4.1. Variables

We are interested in how native speakers of Russian acquire French pronunciation. Our goal is to compare the same target words in French for two cohorts of participants: one *French-native* and one *Russian-native/L2-French learner* at the time of the experiment. we propose to assess the impact of the variables listed in Table 1.

[2]which,incidentally, reduces embedding anisotropy (Ethayarajh, 2019; Timkey and van Schijndel, 2021).

Experiment	Spk	L1	
Modality	=spk	N/A	
	≠spk	=L1	*Fr* *Ru*
		≠L1	*Fr vs. Ru*

Table 1: Modalities of the comparison experiments. spk = speaker ID ; L1 = speaker's mother tongue.

For the `spk` experiment, we want to show if occurrences by the same speaker (=spk) yield lower alignment cost than for different speakers (≠spk), when calculated in the embeddings space. For the `L1` experiment, we verify embedding alignment performance between utterances of two native speakers (=L1), compared to the performance of the alignment between native-speech and L2-speech (≠L1). We are interested in checking how these hypotheses verify and whether they are modulated by model choice.

4.2. Data

The corpus used in all our experiments was collected to investigate the degree to which Russian-native L2 learners of French rely on the phonemes of their native language when pronouncing French words, as well as how this usage correlates with their L2 proficiency level as defined by the CEFR (Common European Framework of Reference for Languages). It consists of recordings of 12 target words within a frame sentence. The detail of the order in which the stimuli are presented is given in Table 2. Frame sentence is: "dis `target_word`" trois fois and distractor sentence is "La roue sur la rue roule, la rue sous la roue reste".

Order Word	1	2	3	4	5	6
tsarine	13	15	32	50	57	78
j'en chie	1	18	28	49	54	72
sérieux	2	23	31	38	51	69
cache-cache	3	14	26	41	58	67
hier	4	21	30	43	63	76
divan	5	20	29	45	52	68
pour Gabriel	6	22	33	48	60	75
louche	7	17	27	40	62	65
tulle	9	16	34	44	53	77
juxtaposer	10	36	46	61	70	73
pas ceux	11	19	25	39	56	74
garage	12	35	47	55	66	71
distractor	8	24	37	42	59	64

Table 2: Speaker metadata for the Russian – French language interference experiment.

There are nineteen participants, among which

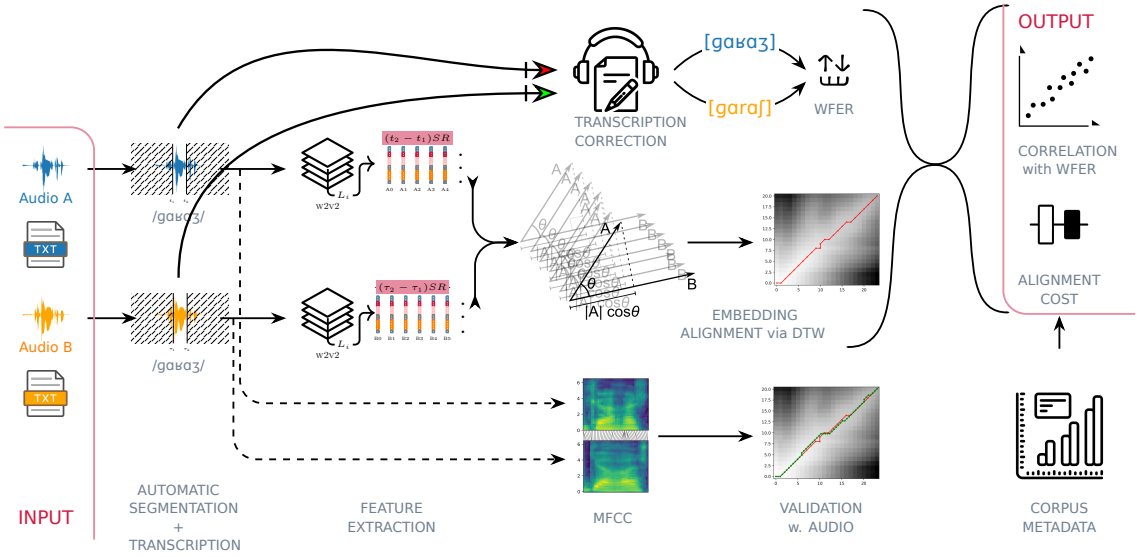

Figure 1: Experimental setup.

ten produce Russian-native accented French (L2 speech) and nine produce native French (native speech). Their age range from 18 to 41 y.o., L2 proficiency from A1 to C2. They were instructed to read out loud 78 randomized sentences, comprising 6 occurrences of the 12 target words and 6 occurrences of a distractor sentence. All recordings were made in the same room in April – May 2019 at the Moscow Lomonosov State University, and all speakers signed an informed consent upon participating to the experiment. Table 3 provides grouped statistics about the participants: natives (All-Fr), learners (All-Ru) including beginners (L2 A1-A2) and advanced (L2 C1-C2), etc. The data is made available in open-access (CC BY-NC-ND-SA 4.0) for verification and reuse in . The high number of occurrences for the same speaker, the shared recording conditions and the available metadata provide a unique framework for speaker-variability studies.

Subgrp	Nb	L1	Age	Gnd	L_{Fr}
All-Fr	9	Fr	23.4	2F;7M	L1
All-Ru	10	Ru	27.5	8F;2M	A1 to C2
L2 A1-A2	4	Ru	25.3	3F;1M	A1-A2
L2 C1-C2	4	Ru	28.5	3F;1M	C1-C2
Fr_ctrl	4	Fr	23.0	2F;2M	L1

Table 3: Average speaker metadata.

The corpus is aligned at word level using Montreal Forced Aligner (MFA. See McAuliffe et al., 2017) on the stimuli list provided to the participants. Alignments were checked manually and corrected by an experienced linguist. The corpus is 38 min 11 s long.

4.3. Experimental setup

The goal of the experiment is to compare different occurrences of a recorded word by measuring the difference in DTW alignment costs after `wav2vec2` feature extraction. These costs are then compared to the error rates determined by an experienced linguist, to verify whether or not the linguist's assessment is correlated with the alignment cost calculated automatically. An overview of the experimental method is presented in Figure 1. All similar pairs of recordings (i.e., same word, different occurrence or speaker, identified with the mention *Audio A* and *Audio B*) are processed using dynamic programming (Giorgino, 2009), either on the embeddings (our main study) or with the audio directly (dashed lines, for verification purposes).

As mentioned in section 4.2, the MFA-generated *phones* tier of each textgrid has been manually corrected to account for the surface realisations of the corresponding recording. These surface transcriptions provide a "gold standard" to our analysis, to which the alignment results can be compared. The phonetic transcriptions are in IPA, and for comparing the transcriptions of two occurrences, the panPhon Normalised Weighted Feature Edit Distance is used (Mortensen et al., 2016). Therefore, instead of the typical Character Error Rate, which accounts for differences without consideration for the phonetic distance between the phonemes replaced, the Weighted Feature Error Rate (WFER) has different multiplication factors according to the nature of the feature differences between reference phoneme and attested phoneme (e.g., 0.25 for a delta on a *velaric* feature, 1 for a *syllabic* feature). In this approach the feature differences between two characters and the weight factors are therefore

syl	son	cons	cont	delrel	lat	nas	strid	voi	sg	cg	ant	cor	distr	lab	hi	lo	back	round	tense	long	velaric
1	1	1	0.5	0.25	0.25	0.25	0.125	0.125	0.125	0.125	0.25	0.25	0.125	0.25	0.25	0.25	0.25	0.25	0.25	0.125	0.25

Table 4: Weights used in Panphon's `jt_weighted_feature_edit_distance`

vectors with one component per phonetic feature.

`jt_weighted_feature_edit_distance` is retained from PanPhon in order to be able to adjust the weight of insertions (i_i) and deletions (d_i) to 0.25 instead of 1. This is done because we consider that read-aloud speech does not favour insertion/deletion errors and they should not bear a weight substantially different from the replaced phonemes since in all likelihood an insertion/deletion will have coarticulatory characteristics in read speech. The weights are stored in a vector $\vec{w_i}$ with the components representing a feature, as illustrated in Table 4. Normalisation is performed by dividing the edit distance obtained by the longest word length. Let A and B denote the two words of length $len(A)$ and $len(B)$ to be compared via WFER :

$$WFER_{A,B} = \frac{\sum_{i=1}^{n_{sub}} \vec{w_i} \cdot \vec{s_i} + 0.25 \cdot (\sum_{i=1}^{n_{del}} d_i + \sum_{i=1}^{n_{ins}} i_i)}{max(len(A),len(B))}$$

with n_{sub}, n_{ins} and n_{del} the minimum number of substitutions, insertions, deletions necessary to turn A into B. Here, $WFER_{A,B}$ represents the phonetic deviation between expert transcription A and B.

Subsequently, the effect of time is analysed separately from the nature of the phonemes encountered by normalising time instead of normalising the embeddings: for a given target word, for all speakers, time is normalized to the mean target word duration value by applying – pre-DTW – a multiplying coefficient to compress or expand the time scale. By doing so, we hope to reduce the effect of time for recordings with large duration differences. The comparison to non time-normalised results shall allow us to determine separately the effect of accuracy and fluency factors in L2-speech.

5. Results

A preliminary study of layer effect is performed, based on the maximisation of the correlation between audio alignment and embeddings' alignment. The nature of the information contained in the embeddings is then addressed starting from section 5.2. The study continues by addressing successively speaker effect and L2 effect on phonologically similar recordings (Sections 5.3 and 5.4).

5.1. Layer effect

In the feature extraction stage, the choice of the layer is experimentally determined in Figure 2 by selecting the layer(s) which best correlate with audio.

A Pearson Correlation evaluation is calculated between audio and embedding alignment costs. Its evolution with layer depth is assessed, to determine the layer most suited to our task.

`XLSR-53` shows a high correlation plateau from layer 0 to 21, then a sudden degradation, such that the last three layers are completely non-correlated with the audio. While the decrease is brutal for `XLSR-53`, it is more progressive for `wav2vec2-FR-7K-large`, which exhibits a high audio-to-embedding correlation on the first layers. Correlation coefficient then gradually decreases with layer depth for `wav2vec2-FR-7K-large`.

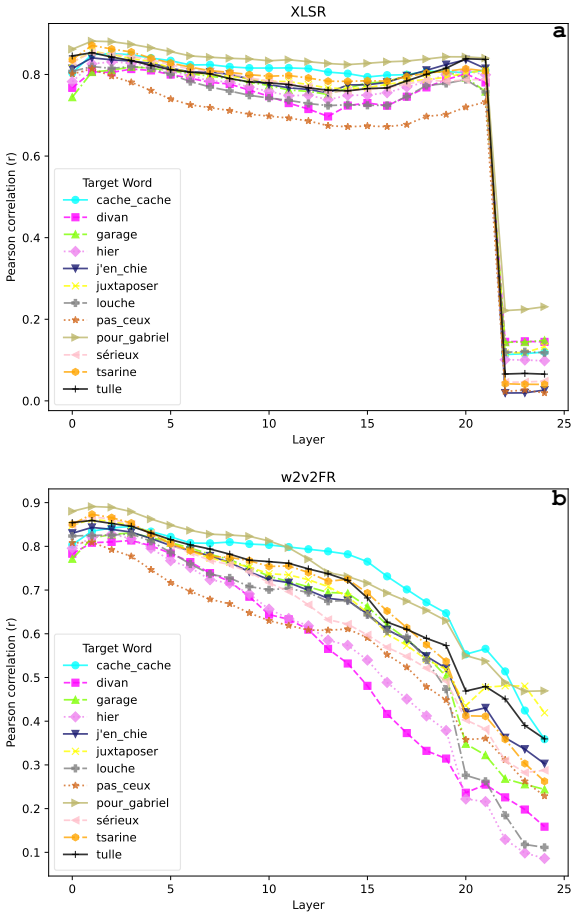

Figure 2: Embedding-based vs Audio-based alignments: Pearson Correlation per Layer ($p<0.05$) for `XLSR-53` (a) and `wav2vec2-FR-7K-large` (b).

This observation motivates the choice of `XLSR-53` layer 16 because audio and embeddings correlate well and the choice coincides with other stud-

ies (Pasad et al., 2021). The optimal layers differ for `wav2vec2-FR-7K-large` model, which suggests that the optimal layer(s) for outputting acoustic/phonetic features are model-specific.

5.2. Effect of the segment

First, the fact that the phonetic information on the segments is present, although self-evident for `wav2vec2` representations (Baevski et al., 2021), is illustrated in Figure 3 by comparing one cosine distance matrix between two different words and between two occurrence of the same word. It shows clearly that the DTW alignment cost on the embeddings of layer 16 is excessively high for words that are different (Figure 3a) when compared to different occurrences of the same word (Figure 3b). For the latter Figure, the alignment path is also more linear, which indicates a more successful alignment.

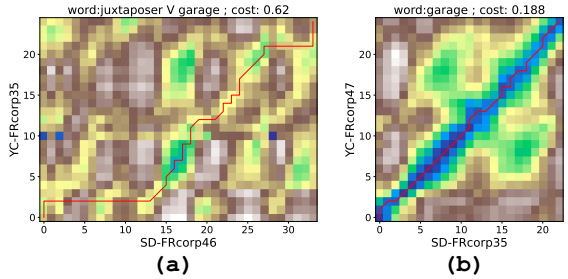

Figure 3: two alignment matrices for two different speakers, (a) compares Fr: "garage" vs. Fr: "juxtaposer" while (b) compares two occurrences of Fr: "garage".

5.3. Speaker effect

Speaker effect is addressed by checking if speaker information is as accessible in the embeddings as in the audio by comparing alignment costs for the `same_speaker` and `different_speaker` modalities. Figure 4 shows the variation of the alignment cost, calculated over all target words for these two modalities and expressed on a custom scale due to the differences in orders of magnitudes. We first give an account of these differences: (1) audio cost and embedding alignment costs just represent the same phenomenon in different units and (2) between z-norm and no-norm: the embedding alignment cost is known to increase after embedding z-normalisation.[3] As for (3) t-normalisation: the fact that the cost is increased after normalising time is counter-intuitive since one would have thought that normalising time would help the recordings match each other better. We show why the increase is homogeneous and perfectly normal. It is due to the way the normalised cost is defined: in our model, we defined the alignment cost for a (n, m) matrix as $\frac{S}{M}$ with S = number of steps to get from the bottom-left to the top-right of the cost matrix and $M = max(m, n)$. on time-normalised embeddings, $n = m$, which artificially maximizes the cost. The differences in numerical cost values mainly stem from how this variable is defined.

We are interested in the ratios between modalities: when shown on a proportional scale, audio cost (4a), z-normalised cost (4b) and non-normalised cost (4c), show relatively identical deltas, which suggests that z-normalisation does not add a significant advantage to the embeddings (z-norm is neither more nor less sensitive to speaker ID). The fact that the `same_speaker` – `different_speaker` alignment cost ratios are the same ratios as the audio is an indication that the acoustic information is present in layer 16, including speaker information.

To go further, Figure 4d shows that t-normalisation tends to reduce the gap between `same_speaker` and `different_speaker` modalities. By normalising time we hope to be able to focus on more abstract phonetic units. We therefore introduce a comparison to the PanPhon (Mortensen et al., 2016) Weighted Feature Error Rate (WFER) which represents in discrete, phonetically-informed terms, the differences at the segmental level between one occurrence of a given word (e.g., [gaʁaʃ]) and another occurrence (e.g., [gaʁaʒ]). Figure 5 shows the alignment cost variation with WFER, calculated over all target words. t-normalisation (Figure 5b) for the `same_speaker` – `different_speaker` configurations, tends to produce wider, thicker violin plot distributions. This results in an increased variance for t-normalised recordings, causing more overlap between the two modalities.

Figure 5a, which focuses on non-normalised embeddings and Figure 5b which treats t-normalised embeddings, are analysed:

- Figure 5a shows that the alignment cost is globally stable with WFER for the `same_speaker` modality and that it increases with WFER for the `different_speaker` modality.
- For Figure 5b, interestingly, we observe the opposite after time normalisation: alignment cost is globally stable with WFER for the `different_speaker` modality and increases with WFER for the `same_speaker` modality.

5.4. Foreign accent effect

L2-speech is analysed in light of the model's performance aligning same occurrences of a word for native speakers or L2 speakers in Figure 6. First of all, the `fr vs. fr` modality exhibits the lowest alignment cost, which was expected.

[3] Same results as in (Le Ferrand et al., 2020).

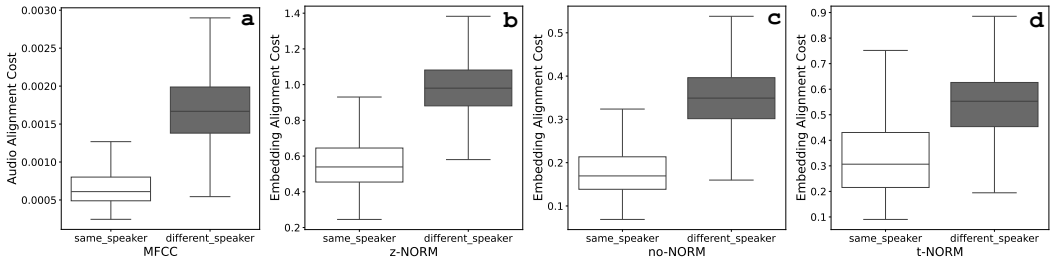

Figure 4: Differences in DTW alignment cost for the `same_speaker` vs. `different_speaker` modalities, for the audio (a), across several normalisation methods: z-normalisation (b) no-normalisation (c) and t-normalisation (d).

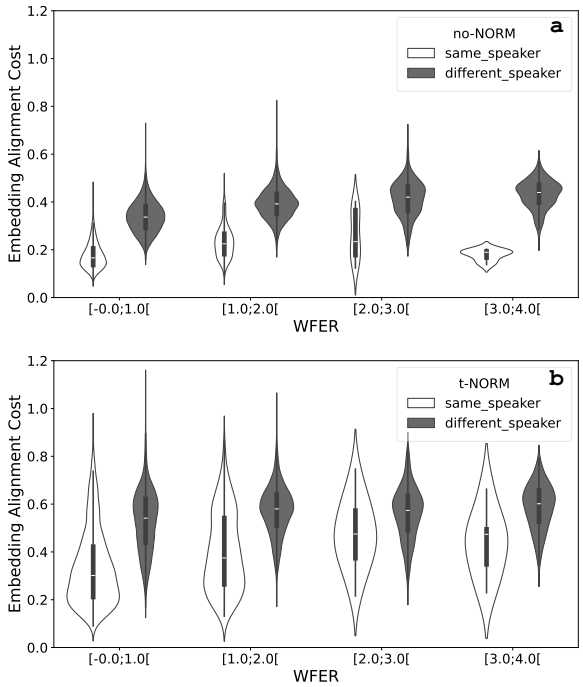

Figure 5: DTW alignment cost variation with WFER for the `same_speaker` vs. `different_speaker` modalities.

By contrast, Figure 6 also shows an odd result: the `ru vs. ru` modality exhibits a higher alignment cost than `fr vs. ru`. This is unexpected because speakers with the same L1, like natives, are expected to exhibit little variability compared to native speech vs. L2-speech (Xie and Jaeger, 2020). In other words, the `ru vs. ru` should have exhibited alignment costs closer to the `fr vs. fr` modality than to the `fr vs. ru` modality. This surprising result led us to split the *L2-French learners* into an advanced and a beginner sub-group. As evidenced in Figures 7a to 7d, this unexpectedly high alignment cost for the `ru vs. ru` modality seems to be rather due to advanced learners (7a, 7b) than to beginners (7c, 7e).

As for t-normalisation effect, overall results (Figure 6) show that the t-normalised setting reduces the discrepancies between groups: the difference in alignment cost between natives and non-natives is less marked after t-normalisation. this is still true after splitting natives into advanced and beginner sub-groups.

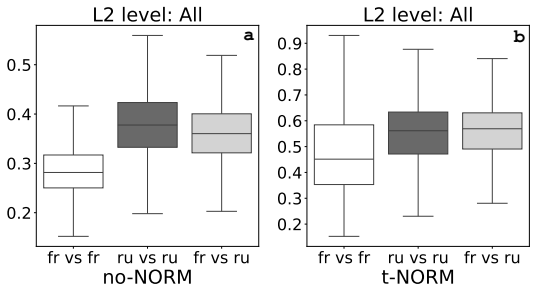

Figure 6: DTW alignment cost for speakers with a different L1: effect of time-normalisation on the whole group of speakers.

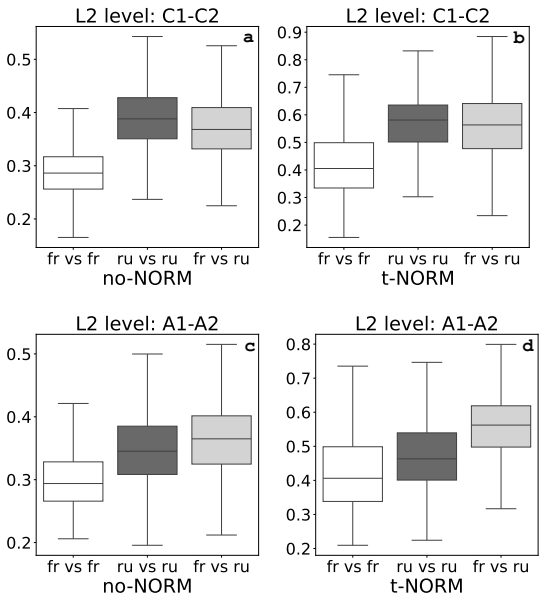

Figure 7: DTW alignment cost for speakers with a different L1: effect of time-normalisation and speakers sub-grouping in levels.

Figure 8 opportunely focuses on the beginner

groups by breaking down the alignment costs shown in Figures 7c and 7d. Zooming in on the `ru vs. ru` modality, non-normalised embeddings have a constant alignment cost. Normalising time proportionally reduces the cost for low WFERs and increases the cost for high WFERs for the `ru vs. ru` modality. This means that differences in speech rate impact more strongly situations with low WFERs than with high WFERs.

Conversely, for the `fr vs. ru` modality, the alignment cost does not vary with WFER in both no-norm and t-norm conditions, suggesting that time is not a dominant parameter for this modality.

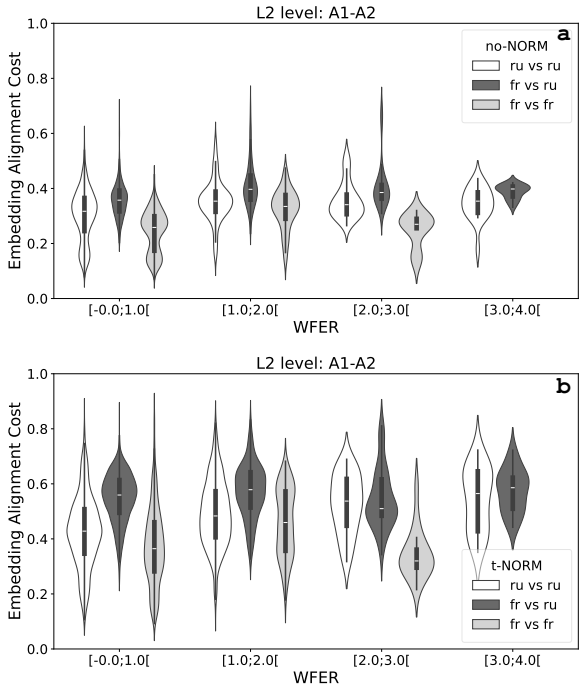

Figure 8: DTW alignment cost variation with WFER for native vs. L2 speech.

5.5. Language Resource References

5.5.1. Online corpus

Our corpus is made up of audio files associated with TextGrids (two per audio file) containing orthographic and phonetic transcriptions, in two versions: the silver version, directly output from MFA, and the gold version, provided by expert phoneticians. It is stored under an Ortolang repository (doi.org/10.82270/ru-fr_interference), is available in open-access with CC BY-NC-SA license (Dashkevich et al., 2026). As a corpus designed for experimental purposes on the acquisition of French (half of the participants) and Russian (the other half), with a single list of participants, a choice of target words phonetically balanced between French and Russian, this resource is valuable for language acquisition studies. Being only half-exploited since the other half (in Russian) did not fit in this submission, the corpus holds potential for future bidirectional studies in French and Russian acquisition. Ortolang is a CLARIN B-Centre (Pierrel et al., 2017).

5.5.2. Online NLP resources

The models used in this study are all available via the huggingface API:

- facebook/wav2vec2-large-xlsr-53
- bofenghuang/asr-wav2vec2-ctc-french

6. Discussion

Our results showed that z-normalization had a very homothetic impact on the embeddings, without any effect on the alignment cost as a function of speaker ID or L1. As a consequence, no z-normalisation was applied to the embeddings.

6.1. Speaker effect

Without time-normalisation, finding any relationship between alignment cost and WFER is a challenge since we were unable to correlate the alignment cost with the Weighted Feature Error Rate. This non-result could be meaningful, in the sense that it could mean that acoustic similarity is more driven by other parameters than the transcriptions, but it would be surprising, especially for a corpus collected in the lab. The literature or foreign accent, by mentioning the *fluency* vs. *accuracy* dichotomy, offered us an angle to approach the issue: comparing t-normalised with non-normalised embeddings.

The fact that t-normalisation reduced the gap between same_speaker alignment costs and different_speakers alignment costs indicates that the time factor, called *fluency factor* in acquisition studies, has for more variability across speakers than within speaker.

This phenomenon, however, is not evenly distributed within the WFER values: for non-normalised embeddings, the alignment costs are closer for low WFERs than for high WFERs, while for t-normalised embeddings, the alignment costs are closer for high WFERs than for low WFERs. It is probable that the effect of t-normalisation is proportionally higher for low WFER values whereas when WFER increases, the variation of the alignment cost is more and driven by segmental differences.

6.2. Foreign accent effect

We first report on a discrepancy between advanced learners and beginners: the beginners seemed more stable in their productions than the advanced learners. Given the very little time alloted to pronunciation acquisition, progress is uneven among

learners and may probably diverge with L2-level. It is also more neglected in classes than syntax or semantics which can be done on paper contrary to pronunciation learning.

Given the alignment cost evolution with WFER for the different modalities, `ru vs. ru` and `fr vs. fr` modality are close to each other for low WFERs after t-normalisation. This means that more of the variation within the `ru vs. ru` modality is explained by a fluency factor than for high WFER values. For high WFER values, however, t-normalizing the data does not significantly improve one modality or another.

We see, to conclude, that the speakers' performance can sometimes be better explained decomposed into a *fluency factor*, materialised in the input by delta-to-mean duration value and an *accuracy factor*, materialised by the Feature Error Rate calculated between two utterances. These two effects are not systematically reflected in the sub-groups, which suggests that all sub-groups are not affected in the same way by the fluency factor.

7. Conclusion

This paper had two goals: investigating whether or not the embeddings encompass information other than abstract linguistic content and devising an approach to address basic linguistic problems using the embeddings in a non-supervised manner.

The first main result is that embeddings contain speaker-specific information because cosine distances between same words are smaller for the same speaker than for different speakers. The delta between these two modalities is the same between audio alignment (MFCCs), normalised or non-normalised embeddings. We showed that normalising time before performing DTW tends to bring alignment cost values closer to each other in proportion, which means that speech rate is one important distinguishing factor between speakers in read speech tasks.

The possibility of t-normalising was explored in order to obtain more speaker-independent representations. The goal is only partially reached because we were only able to reduce the discrepancies between `same_speaker` and `different_speaker` modalities, but speakers remain distinct in the embeddings.

One improvement of the approach would consist in being able to link local cost extrema on the alignment curve with the relative position of the error in the WFER calculation. Such a tool, if developed for didactic purposes, would provide learners with a detailed account of what differs in their productions and they would be able to compare to one or several chosen speakers, which is less normative. Using the method for didactic purpose allows creating targeted pronunciation exercises using live-acquired audio data, which provides interactive feedback to learners.

In terms of potential applications outside didactics, fields like endangered languages documentation can use these methods for interdialectal variability studies. Also, sociophonetics, or even pathological speech studies where privacy is a concern can benefit from these developments without any privacy issue since models are not re-trained with this approach.

8. Acknowledgements

We wish to thank the three anonymous reviewers whose very insightful feedback helped improve this paper.

We are also very grateful for the access to the language acquisition corpora, graciously granted by Daria Dashkevich and Ekaterina Biteeva (LPP). This research could not happen without access to this first-hand laboratory data.

This research was partially funded by the following projects, ranked in alphabetical order :

- The DEEPTYPO project, supported by the *Agence Nationale de la Recherche* (ANR-23-CE38-0003-01),
- The DIAGNOSTIC project, supported by the *Agence d'Innovation de Défense* (grant n° 2022 65 007),
- The DIPVAR project, supported by the *Agence Nationale de la Recherche* (ANR-21-CE38-0019),
- The DLS-HN project, supported by the *Agence Nationale de la Recherche* (ANR-23-CE38-0004).

Last but not least, our deepest gratitude goes to those who lend their voice to phonetics : many thanks to the participants, for volunteering their time and effort in our experiments.

9. Bibliographical References

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for model-

ing variation in speech. *Journal of Phonetics*, 92:101137.

Heather Bliss, Jennifer Abel, and Bryan Gick. 2018. Computer-assisted visual articulation feedback in l2 pronunciation instruction: A review. *Journal of Second Language Pronunciation*, 4(1):129–153.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2426–2430. ISCA.

Tracey M Derwing. 2008. 13. curriculum issues in teaching pronunciation to second language learners. In *Phonology and second language acquisition*, pages 347–369. John Benjamins Publishing Company.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2025. "you'll be a nurse, my son!" automatically assessing gender biases in autoregressive language models in french and italian. *Language Resources and Evaluation*, 59(2):1495–1523.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92(1):233–277.

James Emil Flege. 1981. The phonological basis of foreign accent: A hypothesis. *TESOL quarterly*, 15(4):443–455.

Timnit Gebru. 2019. Oxford handbook on AI ethics book chapter on race and gender. *CoRR*, abs/1908.06165.

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31:1–24.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *ComputEL-5*, Dublin, Ireland.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Self-supervised models of speech processing for haitian creole. *BABEL*, 1091(547):544.

Timothy J Hazen, Wade Shen, and Christopher White. 2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426. IEEE.

Sara Kennedy and Pavel Trofimovich. 2017. Pronunciation acquisition. In *The Routledge handbook of instructed second language acquisition*, pages 260–279. Routledge.

Kartik Khandelwal, Preethi Jyothi, Abhijeet Awasthi, and Sunita Sarawagi. 2020. Black-box adaptation of asr for accented speech. *arXiv preprint arXiv:2006.13519*.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ekaterina Biteeva Lecocq. 2021. *Complexité et contrôle du geste linguo-palatal sous l'éclairage de sa variabilité. Le cas de la palatalisation en russe. Aspects phonétiques et phonologiques.* Ph.D. thesis, Université Grenoble Alpes.

Boris M Lobanov. 1971. Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Alene Moyer. 2013. *Foreign accent: The phenomenon of non-native speech*. Cambridge University Press.

Murray J Munro. 2008. Foreign accent and speech intelligibility. In *Phonology and second language acquisition*, pages 193–218. John Benjamins Publishing Company.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean, and Frédéric Pierre. 2017. Ortolang: a french infrastructure for open resources and tools for language. In *Linköping Electronic Conference Proceedings*, volume 136, pages 102–112.

Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard. 2020. Neural network based end-to-end query by example spoken term detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1416–1427.

Kazuya Saito, Stuart Webb, Pavel Trofimovich, and Talia Isaacs. 2016. Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, 19(3):597–609.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101. IEEE.

Florian Schiel and Andrea Baumann. 2006. Phondat 1, corpus version 3.

Yui Suzukida. 2021. The contribution of individual differences to l2 pronunciation learning: Insights from research and pedagogical implications. *RELC Journal*, 52(1):48–61.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.

Steven H Weinberger and Stephen A Kunath. 2011. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*, pages 265–281. Brill.

Xin Xie and T Florian Jaeger. 2020. Comparing non-native and native speech: Are l2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5):3322–3347.

Anna Zee, Marc Zee, and Anders Søgaard. 2024. Group fairness in multilingual speech recognition models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226.

Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Interspeech*, pages 3168–3172.

10. Language Resource References

Dashkevich, Daria and Biteeva, Ekaterina and Fily, Maxime. 2026. *ru-fr_interference*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

A. Optional Supplementary Materials: Appendices, Software, and Data

A.1. Software and programms

The whole pipeline for (i) automatic alignment, (ii) feature extraction, normalisation, DTW cost calculation and (iii) statistical post-processings is provided

on our github: https://github.com/maxime-fily/RuFr-interference

A.2. Extra space for ethical considerations and limitations

The models used in this study are diverse: one multinlingual pretrained model and one monolingual fine-tuned model. It would have been better to have a third model (e.g., a monolingual pretrained model).

The limited size of our corpus restricted our ability to explore all aspects of speech variability. Limited size does not invalidate our results, but still, applying our method to larger corpora (Schiel and Baumann, 2006) would certainly help validate and extend our findings. In particular, the gender imbalance between the group of natives vs. non-natives, which we tried to address when we did the per-level sub-grouping, would certainly be less important on a larger corpus.

A.3. Morpheme length effect

We verify in figure 9 that the morpheme length in not a confounding factor for WFER. the distribution of length accros all WFER values confirms that length is not correlated with WFER.

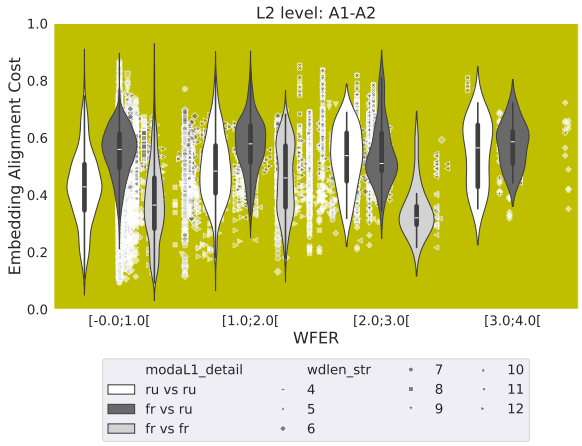

Figure 9: shape-coded length of morphemes for the Foreign accentedness effect: showing that WFER values are not correlated with length.