Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced Alignment with the Italian Component of the ISLE corpus

Nicolas Ballier
LLF & CLILLAC-ARP
Université Paris Cité
rue Thomas Mann
75013 PARIS, FRANCE
nicolas.ballier@u-paris.fr

Adrien Méli
CLILLAC-ARP
Université Paris Cité
rue Thomas Mann
75013 PARIS, FRANCE
adrienmeli@gmail.com

Abstract

Automatic Speech Recognition (ASR) can be used to analyse L2 speech but researchers cannot be sure that the ASR transcriptions accurately represent learner speech. We aim to confront the ASR outputs with the acoustic analysis of learner speech. Whisper (Radford, 2023) provides transcriptions and probabilities associated to the predicted transcriptions. This paper analyses how global phonetic analyses of learner data can be used to potentially confirm these Whisper probability scores assigned to learner transcriptions. We tested the Italian component of the ISLE corpus with phonetic analyses of 23 learners of English. We compared the levels assigned to these speakers by the corpus experts to the outputs of Whisper's `tiny` model. We discuss the phonetic features that may account for these Whisper predictions using acoustic data extracted from forced alignment. We try to correlate the levels assigned to the speakers in the ISLE corpus with the quality of the phonetic realisation, using global vocalic measurements such as the convex hull or Euclidian distances between monophthongs. We show that Levenshtein distance to the reference transcription of the Whisper `tiny` model (measured using Levenshtein distance to the read text) correlates with the grades assigned by the annotators.

1 Introduction

Learner speech has mostly been recently researched with Automatic Speech Recognition (ASR) system and the focus has been on phone substitution (Chanethom and Henderson, 2023). These analyses presuppose time-consuming manual checking of the transcriptions against the recordings. We would like to explore acoustic correlates of ASR transcriptions and investigate whether phonetic data extracted from the transcriptions could be used to confirm the ASR diagnoses. Our Research question is thus: 'Can Whisper's automatically generated transcriptions be used to assess a non-native speaker's pronunciation?' OpenAI's Whisper (Radford et al., 2023) generates time-stamped transcriptions of recorded speech from simple audio files. When mapping the signal to the best candidates for transcription, Whisper ascribes a probability score to each subtoken, which evaluates the likelihood that the transcription that was selected is correct. With non-native speakers, one potential issue is that mispronunciations, especially when systematic or when pertaining to phonemic sequences with dense phonological neighbourhoods, may lead to transcription errors in spite of high probability scores. The purpose of this study is to find out whether vocalic analyses based on force-aligning Whisper's transcriptions provide reliable, usable acoustic information about speakers' characteristics in pronouncing English; 1) to find out whether Whisper's scores correlate with speakers' proficiencies in pronouncing English; 2) to find out whether vocalic data collected from force-aligning Whisper's transcriptions provides reliable information regarding the speakers' performances.

We focus on vowels as they are notoriously difficult (Ballier and Martin, 2015) for learners. We explore several holistic representations of vowels: the acoustic (F1 and F2) formants, the global vowel trapezium plots and the corresponding convex hull as they are likely to be indicative of any actual phonological or phonetic phenomena underlying non-native speakers' pronunciations. Using the recordings of 23 Italian speakers from the ISLE corpus, this study investigates the linguistic significance of Whisper's probability scores, *i.e.,* whether they are indicative of the non-native speakers' proficiencies in pronouncing English. It also explores whether vocalic analyses based

Nicolas Ballier and Adrien Méli. Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced alignment with the Italian Component of the ISLE corpus. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 20–32.

on force-aligning Whisper's transcriptions provide reliable, usable information about the speakers' performances.

Whisper (Radford et al., 2023) is a multilingual audio model trained to do language detection, voice activity detection, transcription translation, and up to a point, diarisation. It was trained on 680k hours of labelled speech data and reports state-of-the-art performance for transcription (Radford et al., 2023). Among these functionalities, the language detection task has not really been used for second language acquisition analysis. The analysis of the probability assigned to the sub-token predicted is still in its early stage (Ballier et al., accepted(a),a). With the C++ implementation of Whisper, we produced the transcriptions and the probability assigned to the subtokens. When accessing the internal representations of Whisper like the probability, linguists do not deal with tokens but with subtokens, which are the results of a byte-pair-encoding process designed to eliminate out-of-vocabulary tokens (Sennrich et al., 2016). This very sub-tokenisation also varies across models, even though Whisper uses the same dictionary of sub-tokens for the different models.

We focused on the Italian component of the ISLE corpus, because the level of the corpus is not homogeneous between the Italian component and the German one. The ISLE corpus derives from a European project aiming at analysing non-native speech, notably English spoken by German and Italian learners. The quality of the English spoken by each speaker was graded (from 1 to 5) and the raters reach a good agreement (Atwell et al., 2003).

The remaining sections of the paper are organised as follows: Section 2 mentions previous research, Section 3 outlines our method, Section 4 delves into our results, and Section 5 provides a discussion of these results.

2 Background Research

Automatic speech recognition (ASR) has been frequently used for the automatic analysis of learner speech (Dalby et al., 1998; Inceoglu et al., 2020; Tejedor-García et al., 2021; Ando and Zhang, 2005), compared to audio Large Language Models (LLMs). The number of papers using Whisper for the investigation of L2 speech is, for the time being, limited, but previous research suggests that the probability assigned to the sub-token can be used as a proxy for the prediction of the levels of the learners (Ballier et al., 2023). Speech recognition is typically used to compute deviations from reference texts in read speech and investigate phone substitutions (McCrocklin et al., 2019; McCrocklin and Edalatishams, 2020; Chanethom and Henderson, 2023). An important contribution is the paper that uses the Otter system to try to measure the shortcomings of the models in relation to the vowel system on a very limited set constraints (Chan et al., 2022). And using the ISLE corpus data, (Arora et al., 2018) try to interpret the mistranscriptions in terms of phonological features, thus focusing more on consonants.

3 Material and methods

In this section, we describe the ISLE corpus data and the pipeline utilised to annotate the data and the main phonetic representations. We analyzed the convex hull representing the trapezium of vowels, the number of vertices produced by the vowel trapezium representation, and then we present the Whisper output.

3.1 The ISLE Corpus Data

The corpus was collected to analyse non-native speech and is available from ELRA. The sections of the ISLE corpus correspond to phonological targets that were tested, with the exception of the read speech task (block A) which contained them all. We re-organised the ELRA data compiled in 1999 in a unique dataset gathering metadata, prompts, objectives and expert annotations. Table 1 illustrates the types of prompts that learners had to read.

The material used in this study comes from the ISLE corpus (Menzel et al., 2000). The recordings of 23 Italian speakers reading 180 blocks of text were analyzed in the fashion described in the following paragraph. The ISLE corpus is particularly interesting to study as it provides standardised recordings of a sizeable sample of speakers, whose performances were evaluated by trained annotators. These features make it possible to obtain two baselines, the script to read and the human evaluations, against which Whisper's performances can be compared.

Block	# Sents.	Linguistic Issue	Exercise Type	Examples
A B C	27 33 22	Wide vocabulary coverage (410)	Adaptation/ Reading	"In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred metres from the top before they had to turn back."
D	81	Problem phones Weak Forms	Minimal Pair Item selection/ combination	"I said bad not bed." "She's wearing a brown wooly hat and a red scarf."
E	63	Stress Weak Forms Problem Phones Consonant clusters	Reading	"The convict expressed anger at the sentence." "The jury took two days to convict him."
F	10	Weak Forms Problem Phones	Description/ Item selection/ combination	"I would like chicken with fried potatoes, broccoli, peas and a glass of water."
G	11	Weak Forms Problem Phones	Item selection/ Combination	"This year I'd like to visit Rome.

Table 1: Typology of prompts in the ISLE data (after Menzel et al., 2000)

3.2 Whisper outputs

We have used the C++ implementation (Gerganov, 2023) of Whisper and the `tiny` model, more likely to be sensitive to non-native deviations from the training model realisations (Ballier et al., 2023). Whisper transcribes speech and the C++ implementation also allows researchers to extract the probability assigned by the Whisper model to the predicted subtokens. Figure 1 gives an example of the probabilities assigned to the predicted subtoken. The lowest probability score here corresponds to a mispronunciation of learner #134 who realises "*weather*" with a long vowel [wi:]. As this example shows, "*weather*" is transcribed as "weeder" in the transcription but corresponds internally to two subtokens (*we—eder*) in the Whisper representations. It is very difficult to re-align subtokens (we—eder) to tokens transcribed by Whisper (weeder) and to map these outputs to the reference ("*weather*"), so that we did not exploit probabilities at the subtoken level but only globally. When modelling data, we only considered the mean value of Whisper's probability scores as a unique datapoint per speaker).

```
[_TT_460]   0.747888
The      0.992373
second   0.995847
difficulty 0.996018
about     0.956371
climbing 0.998327
Everest     0.962417
is         0.991093
the         0.986653
we         0.332225
eder     0.876064
.         0.970952
```

Figure 1: Example of the C++ Whisper output. The subtokens of the Whisper transcriptions are associated to a probability. *[_TT_460]* is a special subtoken corresponding to temporal value. The mistranscribed "*weeder*" (corresponding to "*weather*") is split into two subtokens *we—eder*. The realisation of the first syllable by the learner is phonetically [wi].

3.3 Whisper Scoring

We extracted the probability assigned to each subtoken and to the language assigned by the language prediction functionality. Whisper's probability scores are generated in a file with a subtoken and a score per line. Subtokens often correspond to words, but are sometimes made up of syllables, silences, or punctuation marks such as commas or periods. "*Expedition*", for instance, constitutes a token, but its plural, "*expeditions*", is split into "*exped*" and "*itions*", each with their

respective probability scores. Unfortunately, this feature makes it non-trivial to match the scores with the alignment, so that per-speaker probability scores were simply calculated by averaging over each token's score. Figure 2 shows a visualisation of the different levels of probability assigned to the subtokens by the tiny model. A transcription like "wee—der" (corresponding to "*weather*") shows low probabilities that are consistent with misrealisations of the vowel quality and of the interdental fricative.

3.4 Data Processing

For each speaker, the original short sound files were concatenated into a main audio file and input into Whisper, which in turn generated time-stamped `.srt` subtitles and a `.txt` file listing the probability scores for each token. The time-stamps from the subtitles were then used to split the main audio files into short ones. These short audio files and their matching Whisper transcription from the subtitles were fed into forced-aligner P2FA (Yuan and Liberman, 2008), which generated Praat (Boersma and Weenink, 2019) TextGrids with alignments at the segmental and lexical levels. The reason underlying this seemingly tedious procedure is the contention that feeding the forced aligner with short audio recordings will prevent cascading alignment errors. A syllabic tier and another segmental tier based on the British pronunciations of the *Longman Pronunciation Dictionary* (Wells, 2000) were also added. Finally, all these short TextGrids were merged into one main TextGrid. Vocalic data was then collected by parsing the LPD segmental tier of each speaker's main TextGrid and storing relevant information, such as formant values and duration, when the segment was a vowel.

Figure 3 recaps the different alignments produced with our pipeline. We used the P2FA aligner to process the recordings. The aligner is fed with the CMU phonetic dictionary, one of the rare open source available for English, but which assumes an American pronunciation. We then used the PEASYV pipeline (Méli and Ballier, 2023) to generate the reports on the phone inventories of the different learners. Figure 7 sums up the vowel inventories corresponding to the transcriptions, with the proviso that some learners misread sentences or that some sentences for some speakers are not actually present in the ELRA data.

3.5 Evaluation Metrics

We wanted to correlate the mean probability scoring assigned by Whisper, the grades assigned by the annotators of the corpus (ranging from 1 to 5 for the 23 Italian speakers) and acoustic properties extracted from our forced alignment of the learner recordings with the Whisper transcription.

3.5.1 Levenshtein distance

One metric instrumental to this is study is the Levenshtein distance, which calculates the number of edits needed to change one string of characters into another. The systematic comparison of each speaker's Whisper-generated transcription with the original ISLE script to read, provided by the designers of the corpus, was made after taking the following steps: the script to read was stripped of capital letters, blank spaces and punctuation marks. Measurements written in full letters were converted to numbers, in keeping with Whisper, which transcribes most numbers in Arabic. Subtleties such as "3 meters", transcribed by Whisper in Arabic numbers, but "three mountains" transcribed in full letters, were accounted for. Each speaker's Whisper-generated transcription underwent the same treatment: blank spaces and punctuation marks were removed, and capital letters were converted to lower-case.

3.5.2 Main Acoustic Correlates

The next step was to determine whether correlations existed between the two baselines of the ISLE corpus, *i.e.* the annotators' grades (from 1 to 4) and the Levenshtein distance to the original script (formatted in the way described in the previous section). In order to do so, several phonetic metrics for each speaker were computed with the formant values extracted at mid-temporal values from our forced alignment:

1. the Euclidean distances of each monophthong to all other 11 monophthongs in the F1/F2 space using mid-temporal values (66 datapoints per speaker);

2. the Vowel Inherent Spectral Change (Nearey and Assmann, 1986; Nearey, 2012; Morrison and Nearey, 2007; Morrison, 2012) of each vowel, *i.e.,* both monophtongs and diphthongs, using the mean formant values at 20% and 80% of the vowels' durations (19 datapoints per speaker);

There are three main difficulties facing any part in attempting to climb Everest.
The first of these difficulties is that of altitude.
At great age the air is very thin.
The air contains so little oxygen that climbers can only move very slowly.
They also think more slowly and these make-outs then to make mistakes.
There is one way of reducing this difficulty.
If I climb Everest 6,000 meters or so and stays at the altitude for a few days,
he will become a user to the fin air.
This process of getting user to the altitude is called acclimatization.
However, once the climber reaches an altitude of 8,000 meters, acclimatization is not longer possible.
Instead of acclimatization, the body suffers damage.
The musculus becomes smaller and loads their straight and within a few days the climber is
no longer able to move. The summit of Everest is over 8,800 meters.
It would be best to climb the last 2,500 meters in only one or two days but it is not possible.
A second way of reducing the difficulty of altitude is for each climber to carry oxygen in a bottle
on this bag. This oxygen equipment we knew from earlier expedition must be light in weight.
The second difficulty about climbing Everest is the weeder.

Figure 2: Probability Scoring of Whisper's Tiny model predictions for the subtokens of the transcription of (male) speaker 134. Purple corresponds to high probability, cyan to low probability

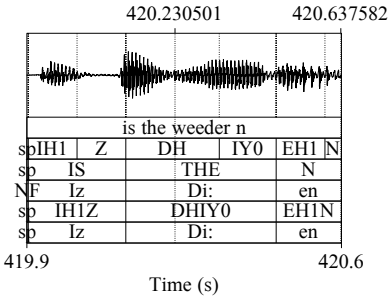

Figure 3: Fragment of a TextGrid corresponding to "*is the weather n*". Under the waveform are the five tiers that correspond to the Whisper output transcript ("is the weeder n"), to the phoneme (CMU Arpabet transcription), to the words ("weeder" was missed by the aligner), to the British transcription (SAMPA), to the syllables of the CMU Arpabet transcription and to the British transcription (SAMPA) of the syllables.

3. the area of the speaker's convex hull and its number of vertices (2 datapoints).

Pearson correlations with the Levenshtein distance and the annotators' marks were then systematically computed. p-values above the 0.05 conventional threshold were rejected, along with absolute R-values inferior to 0.55, in order to exclude weak correlations.

3.5.3 Probability Density and Kernel Density Estimates (KDE), Convex Hulls and Number of Vertices

We wanted to test several global validation procedures based on acoustic correlates of vowels, investigating the convex hull and number of vertices as a representation of the trapezium produced by the different learners as compared to potential British models (the pronunciation norm indicated for the ISLE data). We used the British pronunciation norm, reported as being the one used by the learners in the corpus (Atwell et al., 2003). We computed the convex hull and the number of vertices needed to represent the trapezium of vowels for the speakers. Figure 11 illustrates the trapezoids of the Italian male and female speakers, the vertices connecting the means of the F1/F2 values for vowels. The reference trapezium corresponds to the values reported in one of the reference studies for British English (Deterding, 1997). Because the formant extractions were based on lab speech (vowels in the /hVd/ context), these means correspond to hyperarticulated values. Our last attempt at exploiting the area of the vowel space is the number of vertices associated to the different vowel trapezia representing the vowel plots. The mean of each vowel distribution serves as an edge for the vowel space trapezoid and we reported the number of vertices. The hypothesis in terms of the number of vertices was "the higher the number of vertices, the bigger the vocalic space", and then the clearer or the more separate the various vocalic realisations are and therefore the better the overall pronunciation might be.

4 Results

In this section, we present the different results from individual realisations of vowels to more global comparisons.

4.1 KDE of Vowel Realisations

We used kernel density estimates (KDE) to represent in three dimensions, the F1 and F2 probability density. We are using this visual representation as a cue for the separability of the different

vowels. We would expect the properly realised phonetic minimal pairs to be realised as two distinct cones. Conversely, when only one vortex or pyramid can be observed, we assume that the distinction between the two phonemes is not realised. We computed these KDE for the vowels shown in Figure 7, and we only show here the most relevant pair of confusing vowels (KIT vs. FLEECE) illustrated by two speakers for our learners.

4.2 Number of Vertices

Table 2 reports the number of vertices that is associated with the different vowel trapezia representing the vowel plots. Our hypothesis was "the higher the number of vertices, the bigger the vocalic space" and then the clearer or the more separate the various vocalic realisations were and therefore the better the overall pronunciation might be. This hypothesis is not verified, at least with our data.

Level	mean of vertices	support
1	6.71	7
2	6.73	11
3	6.50	4
4	6.00	1

Table 2: Mean of complex hull vertices per level for Italian speakers

4.3 Reference Vowel Inventories

For a global analysis, we tried to come up with a reference inventory of the phoneme systems, the vocalic system, because most of the subjects were assumed to have British pronunciation. We used the British transcription from the Longman Pronouncing Dictionary to try to estimate the reference vowel inventory. Such an undertaking is challenging because we need to eliminate the variants that are automatically assigned by the phonetic aligner. The variants, when available in the dictionary of the aligner, are selected on the basis of the acoustic signal. We systematically took the first variant when several were present.

The distribution of the vowel inventories that we would expect varies across speakers but we do not report phoneme error rates, we are trying to offer a global appreciation. This is based on the transcription of the target, the text that needed to be read by the different learners following the different tasks of the ISLE data. A total of 30,032 vowels across the 23 Italian speakers were collected and analyzed. No filters, such as removing function words or focusing on stressed syllables only, were applied. The per-vowel distributions can be found in Figure 7. Monophthongs account for 79% of all collected phonemes, with /ə/ amounting to 19.2% of all vocalic occurrences with 5,757 tokens.

4.4 Correlations to Levenshtein Distance

The Levensthein distance to the reference text read by the ISLE speakers (the smaller the distance, the better the pronunciation) proved to be robustly correlated to per-speaker mean of the probability scores (R=-0.94), to the ability to classify monophthongs (R=-0.7), and partially correlated to the learner grades (R=-0.57) assigned by ISLE annotators.

The main result is a strong correlation ($R = -0.94, p < 0.005$) of Whisper's probability scores with the Levenshtein distance separating the transcriptions from the script of the reading assignment. Figure 8 confirms the hypothesis that higher probability scores in the Whisper prediction corresponds to a better pronunciation (lesser deviation from the expected realisations). Speakers whose automatic transcriptions have a higher Levenshtein distance are more likely to have lower Whisper probability scores.

The second observed correlation is with the acoustic data. The Levenshtein distance is partially correlated to the ability to classify monophthongs for each speaker on the basis of their formant values. We extracted the formant values from the forced aligned data and used the k nearest neighbour (k-NN) algorithm (Deng et al., 2016) to classify the monophthongs on the basis of their F1/F2 formant values[1]. We computed the accuracy for this classification task. The scatter plot on Figure 9 illustrates the relationship between the Levenshtein distance to the original text string and the accuracy reported for the per-speaker classification of Italian speaker's monophthongs using the k-NN algorithm (k-NN Accuracy) represented on the y-axis. One can see a clear negative correlation between the Levenshtein distance and this k-NN Accuracy, as indicated by the downward sloping trend line and Pearson's correlation coefficient R = -0.7. The relationship is statistically significant, with $p < 0.001$. The data points are some-

[1]Vowel discrimination between native and non-native realisations have already been tested with this type of clustering (Méli and Ballier, 2019).

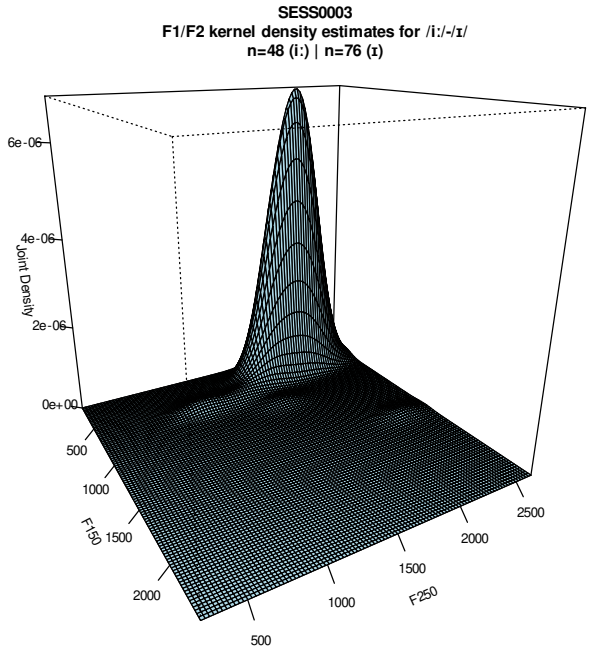

Figure 4: KDE estimate for F1 / F2 probability density for Speaker #S003

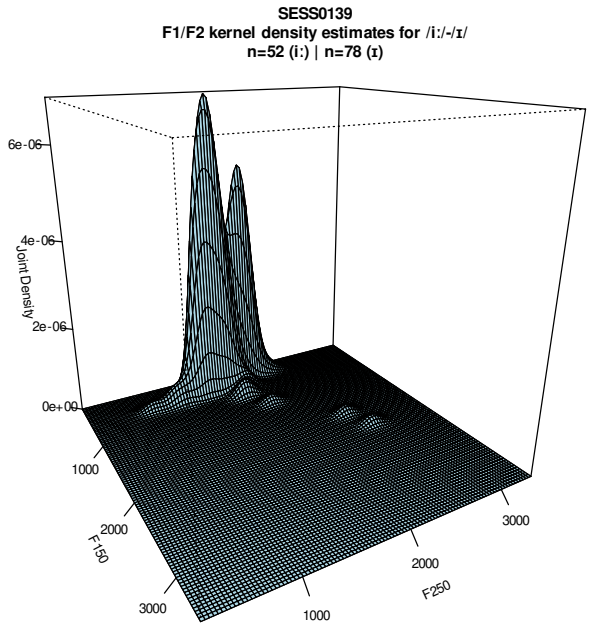

Figure 5: KDE estimate for F1 / F2 probability density for Speaker #139

Figure 6: Comparison of the two Kernel Density Estimates (KDE) for the KIT vs. FLEECE vowels for two speakers. The unimodal distribution of the acoustic realisations (one peak) suggests that speaker #3 does not properly categorise the two vowels (top), whereas speaker #139 produces two distinct series of realisations (bottom) for the KIT vs. FLEECE vowels, suggesting that the vowel categorisation has been properly acquired.

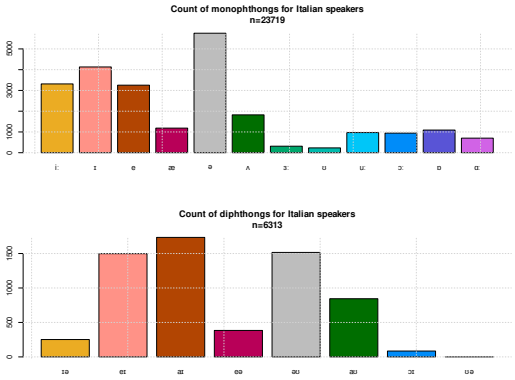

Figure 7: Vowel inventories aggregated on the 23 Italian Speakers, monophthongs (top) and diphthongs (bottom), based on the forced alignment of the tiny Whisper transcriptions

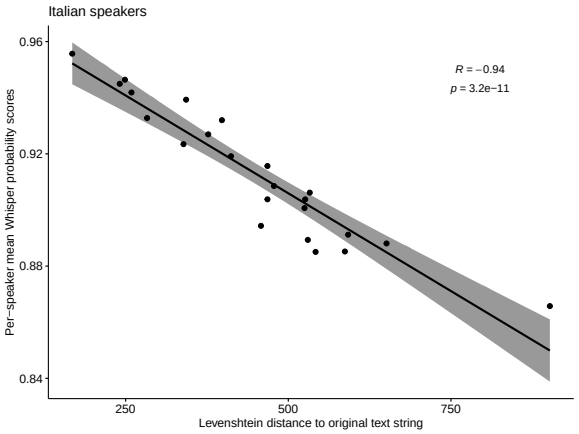

Figure 8: Negative correlation between the Levenshtein distance to the original text string and the per-speaker mean of the Whisper probability scores. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

what scattered around the trend line, but generally follow the negative trend. The grey shaded area represents the confidence interval around the regression line, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points. This plot confirms that our use of the Levenshtein distance is a sensible correlate to the assessment of phonetic quality: the more a text is altered from its original form, the harder it becomes for the k-NN algorithm to accurately classify the monopthongs of a given speaker. Admittedly, the accuracy reported is far from perfect, as the accuracy of the prediction (with 70% train, 30% test) ranged from about 0.35 to 0.55, but it should be borne in mind that vowel data points for monophthongs partially overlap, so that accuracy for native speakers' monophthong classification would also be limited. With a skewed distribution and 12 classes to predict, this is no easy task. Nevertheless, this accuracy of the classification of the monophthongs on the basis of their formant values correlates with the Levenshtein distance.

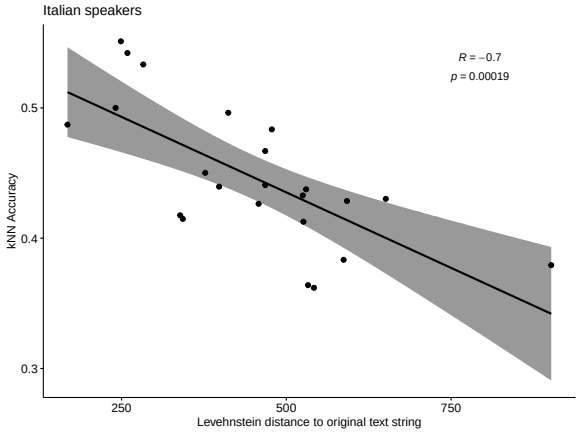

Figure 9: Negative correlation between the Levenshtein distance and the accuracy of the prediction of the monophthongs using k-NN. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

Finally, the correlation of the Levenshtein distance with the grades assigned by the ISLE annotators are weaker but the correlation remains statistically significant ($R = -0.57$, $p < 0.005$). Figure 10 suggests that as the Levenshtein distance increases (indicating greater difference from the original text), the annotators' marks tend to decrease. This means that the annotators' grading of the Italian speakers does decrease when the Whisper `tiny` model transcriptions deviate more from the original text. The Levensthein distance is therefore a metric consistent with the grades assigned to the Italian learners in the ISLE corpus.

4.5 Absence of Global Correlations

However, the analyses of 88 parameters related to vocalic data (*e.g.,* the Euclidean distances between each monophthong in the F1/F2 vocalic space) return no, or very weak, correlations with the Levenshtein distance. One exception may be found in the /iː/-/ɪ/ distance ($R = -0.56$, $p < 0.005$). This validates our hypothesis that visual inspection of the KDE density of these two vowels is

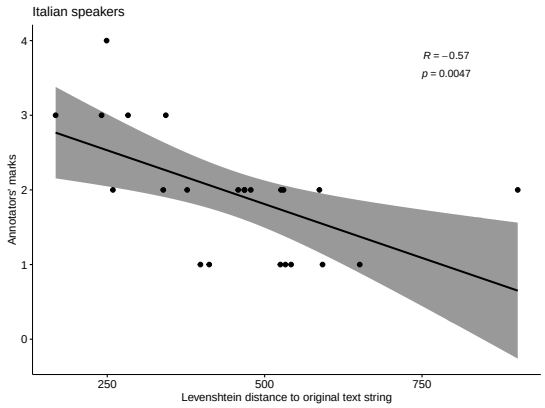

Figure 10: Negative correlation between the Levenshtein distance and the corpus annotator's grades. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

a valid cue for the interpretation of the quality of realisations. This distinction between the two phonemes is noted in the papers describing the ISLE corpus (Atwell et al., 2003) and in the L2 phonetic literature (Kenworthy, 1987). This global representation of the probability density for F1 and F2 for these two vowels show distinct visual representations. We assume that phonetic realisations are distinctive if two peaks can be distinguished. Conversely, learners failing to mark F1/F2 differences for these vowels have a unimodal distribution. As can be seen on Figure 6, speaker 003 has a unimodal distribution for the F1/F2 realisations of the FLEECE and KIT vowels.

5 Discussion

To the best of our knowledge, this is the first paper that tries to correlate the grades assigned to taylor-made spoken corpora to Whisper outputs (transcriptions and internal representations of their probabilities) and phonetic correlates extracted from forced alignment of the Whisper transcriptions. Assuming we take the ISLE grades as golden reference taken for granted, the discussion bears on how we collected the phonetic data points (subsection 5.1), aggregated the Levenshtein distance neglecting task effects (subsection 5.2), compared scores of linguistic units varying in size and scope (subsection 5.3), measured the correlations (subsection 5.4) and on the Whisper outputs we have not investigated yet (subsection 5.5).

5.1 Precision of the Aligners

The first point to discuss is the precision of the aligner, the tool that automatically aligns the Whisper transcription to the signal. As shown in Figure 3, there may be errors in the forced alignment. We have used the P2FA aligner whose performances may be lower than more recent ones. The Montreal Forced aligner (McAuliffe et al., 2017) may produce better results, but is not that easy to integrate into our annotating pipeline. Previous research suggests that the precision of our pipeline may be lower than more recent ones (Méli et al., 2023). One key question is therefore that of the accuracy of the forced alignment performed by P2FA. A hopefully convincing way to answer this question is to plot the means of the mid-temporal F1 and F2 formant values and to compare them to established values in the literature. Figure 11 shows that the vocalic trapezoids thus obtained for per-sex average values are consistent with those listed in Deterding (1997). The lines trace the convex hulls of the sets of average F1/F2 values. Unfortunately, the number of vertices required to represent the trapezium did not present a consistent pattern correlating with Levenshtein distance or learner grades.

5.2 Task Effects: the Different Prompts

We merged the different sound files corresponding to the subtasks (see Table 1) to analyze the ISLE data and reported global results, in line with the global grading of the sound files by the annotators. We do not report the probability scores or the Levenshtein distance per type of prompts (see Table 1) and do not investigate whether some task effects could be measured, looking at the language prediction and the average probability assigned to the subtokens of the different group. A related research question is the need to estimate what would be the optimal duration of the data to be used by automatic systems when predicting the level of the learners.

5.3 Granularity and Scope of Scoring

Our analysis focuses on vowels, but the papers presenting the ISLE corpus also insisted on phone substitutions for consonants (Atwell et al., 2003). One of the difficulties of using Whisper scoring is that probability scores correspond to subtokens, which do not exactly correspond to syllables and rarely match phonemic representations.

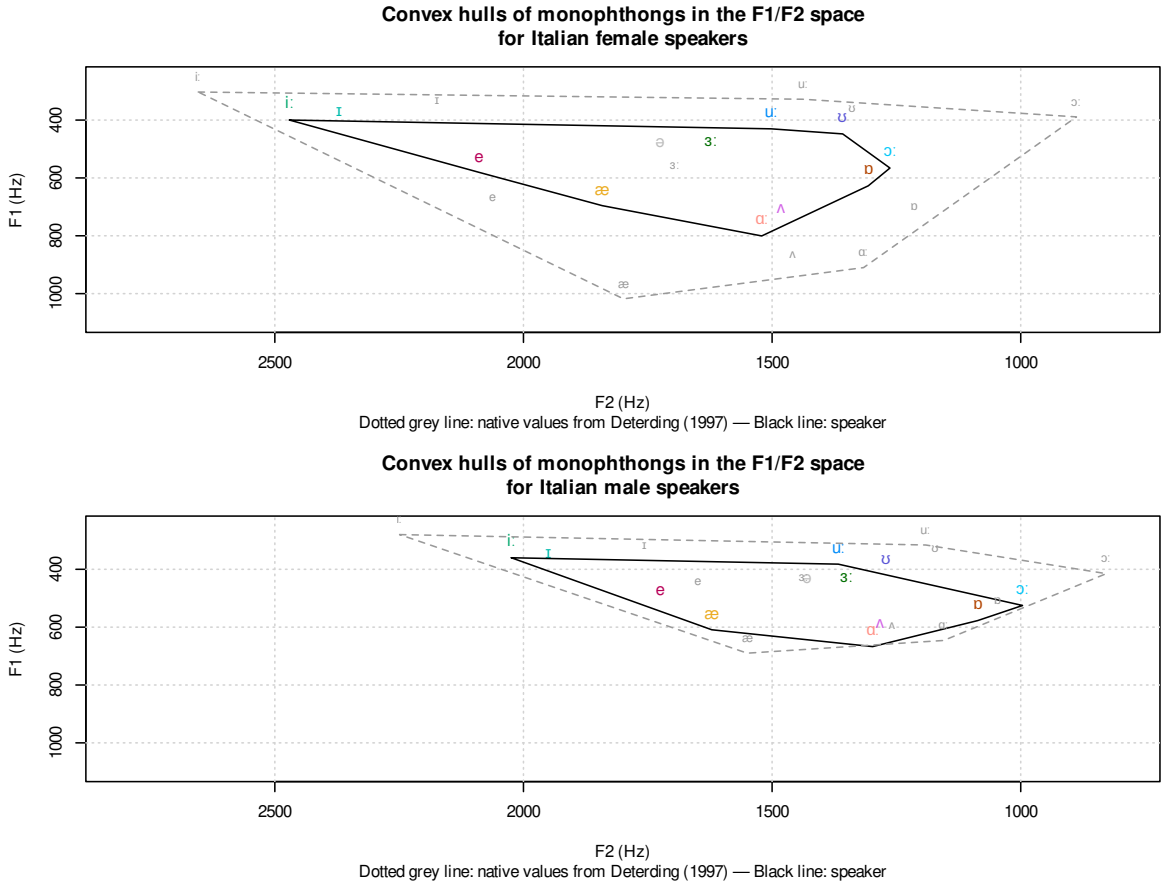

Figure 11: Convex Hulls of female (top, 6 vertices) and male speakers (bottom, 5 vertices) for all the Italian speakers. The dashed trapezium corresponds to the British reference means in lab conditions (hyperarticulation) reported in (Deterding, 1997)

We could at best report the confidence of the system (the Whisper probabilities) at the subtoken level, but in this paper, we mostly consider holistic estimations: Levenshtein distance (the transcription of the full recording) or mean probability of all the predicted subtokens. These 'textual' Whisper predictions can only be partially mapped to speech units of a different granularity. A phonemic transcription could parallel the Levenshtein distance and include phonological consonants. The PEASYV pipeline extracts formants and focuses on vowels. Vowel plots and their structures (numbers of vertices) are holistic representations, and so are vowel inventories. Most other metrics are between two vowels (distances or kernel density estimates) and may be used to monitor whether vowel distinctions have been acquired.

5.4 Alternative Methods

For our 'kitchen sink method', the phonetic variables related to vowel plots reveal little correlation to the levels assigned by the annotators or to the Levenshtein distance. In particular, the number of vertices (at least based on our forced alignment) does not seem to be a plausible correlate for the level of the learners. An ablation method for a multinomial ordinal regression may highlight other variables. Another approach of the vowel realisations is based on Pillai scores. For an intrinsic measure of the dispersion of L2 speech, we may use Pillai scores applied for L2 speech for vowels (Mairano et al., 2019, 2023). Additionally, we have not explored clustering techniques that would try to investigate if the grouped phonetic datapoints corresponding to the reference grades of the corpus would produce consistent results. Assuming there are actually four levels to be considered for the Italian ISLE speakers, what would be the confusion matrices of these four levels using k-means (with k equals 4 for the four levels) for the vowel acoustic correlates we examined? Would the four clusters produced by the k-means correspond to the four levels of the corpus grades?

5.5 Whisper Scoring of Language Detection

Another feature is worth investigation. Whisper has been trained to recognise the language spoken as an identification task (Radford et al., 2023). This language identification (and its associated probability) could be potentially used to analyse learner data, to discriminate speakers predicted to be English or Italian. For example, using Whisper's `large` model to predict the language spoken by the Italian speaker, we observed that more advanced learners (level 3 or 4) were labelled as English, whereas the learners graded as level 2 were either predicted as being English or Italian. With presumably the most robust Whisper model, there seems to be a threshold between less advanced learners whose first language is predicted (Italian) and more advanced learners that are detected as being English. The most interesting case study is the intermediate group of Italian learners labelled "2" in the ISLE corpus that is sometimes predicted as English or as Italian. We want to analyse the potential phonetic correlation that may account for this judgement, therefore potentially validating the idea of a threshold detecting between less advanced learners and more advanced learners with Whisper. We intend to compare these Whisper predictions with the phonetic realisations (including consonants) using the P2FA aligner to compare the various phonetic realisations with the Whisper predictions, trying to account for that difference in the system.

6 Conclusion

In this paper, we have tried to correlate Whisper's transcriptions with the levels assigned to the Italian learners in the ISLE corpus and with acoustic correlates of vowels. We used the Levenshtein distance to measure deviation from the read texts for each speaker based on Whisper's ASR transcriptions and we used forced alignment and the PEASYV annotation pipeline (Méli and Ballier, 2023) to produce our vowel-based acoustic data (vowel formants), phone reports and phonetic measurements. Levenshtein distance does correlate with the levels, but the acoustic correlates we analysed are not convincing. The assumption that Whisper scoring could be a good cue to the quality of the phonetic realisation is validated because it is negatively correlated to the deviation from the reference read text measured with the Levenshtein distance. Our explorations of the holistic phonetic correlates is less successful. Holistic representations like vowel plots apparently fail to be correlated to the grades attributed to the Italian learners in the ISLE corpus. Nevertheless, the type of trapezoids we produced with the PEASYV pipeline could be used in Computer-Assisted Pronunciation Training (CAPT) systems (Rogerson-Revell, 2021) as actionable visualisations for teachers and expert users.

Limitations

The first limitation is the number of speakers and languages for our analysis. Because graded spoken learner corpora are not that frequent, we focused on the ISLE data, and only on the Italian component, since the German component has a different level distribution. Only 11 male speakers were analysed, which also introduces a gender limitation to our work. A second limitation is the focus on segmental errors, like many studies based on Automatic Speech Recognition. The analysis of L2 speech should also account for suprasegmental features. Last, our metrics, visualisation and k-NN analysis of the vowels mostly tackled monophthongs and not diphthongs and these techniques in investigating vowel separability may be contradicted by perception tests.

Ethics Statement

It is important to note that the Whisper scoring should not be used as a substitute for human feedback. Whisper does not explicitly monitor suprasegmental features. As noted during our analyses, the probabilities associated with the Whisper transcriptions do not necessarily guarantee that the word transcribed by Whisper is the most accurate rendition of what the learner actually pronounced. As a consequence, we endorse a human-in-the loop approach to this kind of technology.

Acknowledgments

We thank the NLP4CALL anonymous reviewers, Sara Ng and Alice Henderson for their comments on a previous version of this paper. We also thank Jean-Baptiste Yunès for his implementation of the extraction of the Whisper probabilities.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition*, 6:1817–1853.

Vipul Arora, Aditi Lahiri, and Henning Reetz. 2018. Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America*, 143(1):98–108.

Eric Atwell, Paul Howarth, and Clive Souter. 2003. The isle corpus: Italian and German spoken learner's english. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27:5–18.

Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. accepted(a). Whisper for l2 speech scoring. *International Journal of Speech Technology*, 27(4):–.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara Ng, and Jean-Baptiste Yunès. accepted(b). Probing Whisper Predictions for French, English and Persian Transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, Online. Association for Computational Linguistics.

Nicolas Ballier and Philippe Martin. 2015. Speech annotation of learner corpora. *The Cambridge handbook of learner corpus research*, pages 107–134.

Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023. Using whisper LLM for automatic phonetic diagnosis of L2 speech, a case study with French learners of English. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 282–292, Online. Association for Computational Linguistics.

Paul Boersma and David Weenink. 2019. Praat: doing phonetics by computer [computer program]. version 6.1.07, retrieved 26 november 2019 from http://www.praat.org/.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Proc. Interspeech 2022*, pages 1273–1277.

Vincent Chanethom and Alice Henderson. 2023. Alignment in ASR and L1 Listeners' Recognition of L2 Learner Speech: French EFL Learners & Dictation.Io. *Research in Language*, 21(3):245–266.

Jonathan Dalby, Diane Kewley-Port, and Roy Sillings. 1998. Language-specific pronunciation training using the hearsay system. In *Speech Technology in Language Learning*, pages 25–28.

Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. 2016. Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143–148.

David Deterding. 1997. The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association*, 27(1-2):47–55.

Georgi Gerganov. 2023. whisper.cpp : A high-performance inference of OpenAI's whisper automatic speech recognition (ASR) model.

S. Inceoglu, Hyojung Lim, and Wen-Hsin Chen. 2020. ASR for EFL pronunciation practice: Segmental development and learners' beliefs. *The Journal of Asia TEFL*, 17(3):824–840.

Joanne Kenworthy. 1987. *Teaching English pronunciation*. Longman.

Paolo Mairano, Caroline Bouzon, Marc Capliez, and Valentina De Iacovo. 2019. Acoustic distances, pillai scores and lda classification scores as metrics of l2 comprehensibility and nativelikeness. In *ICPhS2019*, pages 1104–1108.

Paolo Mairano, Fabián Santiago, and Leonardo Contreras Roa. 2023. Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels. *Languages*, 8(4):280.

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.

Shannon McCrocklin and Idée Edalatishams. 2020. Revisiting popular speech recognition software for ESL speech. *Tesol Quarterly*, 54(4):1086–1097.

Shannon McCrocklin, Abdulsamad Humaidan, and Idée e Edalatishams. 2019. ASR dictation program accuracy: Have current programs improved? *Pronunciation in Second Language Learning and Teaching Proceedings*, 10(1):191–200.

Adrien Méli and Nicolas Ballier. 2019. Analyse de la production de voyelles anglaises par des apprenants francophones, l'acquisition du contraste /ɪ/–/iː/ à la lumière des k-nn. *Anglophonia / Caliban-French Journal of English Linguistics*, 27.

Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association.

Geoffrey Stewart Morrison. 2012. Theories of Vowel Inherent Spectral Change. In *Vowel Inherent Spectral Change*, pages 31–47. Springer Science + Business Media.

Geoffrey Stewart Morrison and Terrance M. Nearey. 2007. Testing theories of vowel inherent spectral change). *J. Acoust. Soc. Am.*, 122(1):EL15–EL22.

Adrien Méli and Nicolas Ballier. 2023. PEASYV: A procedure to obtain phonetic data from subtitled videos. *Proceedings of the International Congress of Phonetic Sciences*, pages 3211 – 3215.

Adrien Méli, Steven Coats, and Nicolas Ballier. 2023. Methods for phonetic scraping of youtube videos. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 244–249.

Terrance M. Nearey. 2012. Vowel Inherent Spectral Change in the Vowels of North American English. In *Vowel Inherent Spectral Change*, pages 49–85. Springer.

Terrance M. Nearey and Peter .F. Assmann. 1986. Modeling the role of vowel inherent spectral change in vowel identification. *JASA*, 125:2387-97.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Pamela M Rogerson-Revell. 2021. Computer-assisted pronunciation training (CAPT): Current issues and future directions. *Relc Journal*, 52(1):189–205.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Cristian Tejedor-García, Valentín Cardeñoso-Payo, and David Escudero-Mancebo. 2021. Automatic speech recognition (ASR) systems applied to pronunciation assessment of L2 Spanish for Japanese speakers. *Applied Sciences*, 11(15):6695.

John C. Wells. 2000. *Longman Pronunciation Dictionary*. Pearson Longman, London.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America,*, 123(5):5687-5690.