Whisper for L2 speech scoring *

Nicolas Ballier[1,2*], Taylor Arnold[3], Adrien Méli[1],
Tori (Thurston) Fullerton[1], Jean-Baptiste Yunès[4]

[1*]CLILLAC-ARP, Université Paris Cité, rue Thomas Mann, Paris,
75013 , France.
[2]LLF, Université Paris Cité, rue Thomas Mann, Paris, 75013, France.
[3]Data Science & Linguistics, University of Richmond, 211 Richmond
Way, Richmond, 23226, VA, USA.
[4]IRIF, Université Paris Cité, rue Thomas Mann, Paris, 75013 , France.

*Corresponding author(s). E-mail(s): nicolas.ballier@u-paris.fr ORCID :
0000-0003-2179-1043;
Contributing authors: tarnold2@richmond.edu ORCID : 0000-0003-0576-
0669; adrienmeli@gmail.com; torithurston2020@gmail.com;
jean-baptiste.yunes@u-paris.fr ORCID: 0000-0003-0291-4263;

Abstract

In this paper, we examine whether confidence scores produced by the C++ re-
implementation of Whisper [1] can be used to score L2 learners of English and
classify them. We test whether the language prediction and its probability can be
used to classify French learners of English using a specifically collected dataset
for read speech and a graded corpus, the ANGLISH corpus [2]. We show that
probability scores associated with the Whisper subtokens can be used to classify
learners into levels using the knn algorithm. We show the limitations of the lan-
guage detection probability beyond an initial threshold where the native language
L1 of the learner can actually be predicted by the speaker. We have also used the
ISLE corpus [3] to test the prediction of the levels of Italian and German learners
of English [4]. We show how language detection for Whisper's multilingual larger
models can be used to detect less advanced learners' first language but cannot be
used for learner level classification with advanced learners. Using a greedy align-
ment algorithm, we also discuss the confidence score assigned to Whisper output

---

subtokens and how this may be used for speaker scoring, prediction of learner levels, and learner feedback. We show that low confidence scores and alternative transcriptions can be used as potential cues for learner pronunciation errors.

Keywords: audio LLM, Whisper, ASR, L2 speech, Computer-assisted pronunciation teaching (CAPT)

1 Introduction

The use of Automatic Speech Recognition (ASR) in pronunciation training dates back to the 1990s. A preliminary study pioneered the use of ASR in L2 pronunciation [5], showing that ASR helped improve intelligibility in the learner's L2 and that the improved targeted phonetic contrasts (/i:/ vs. /ɪ/, /θ/ vs. /s/) were also found in untrained words. [6] compared human and ASR evaluations of speech quality. Some researchers explored ways to integrate ASR in pronunciation training programs [7], while others focused on the creation of feedback derived from the ASR transcriptions. More recent studies [8] used Google's ASR to measure the intelligibility of L2 speech (Taiwanese L1, English L2) and concluded that the rating-agreement between the ASR and native speakers mostly depended on both the individual speakers and the speech style (i.e., word lists, read text or more natural speech). Similar systems have been developed with Open Source release, such as KALDI [9], Vosk [1], wav2vec 2.0 [10]. Previous studies focused on the discrepancies between the ASR output of L2 speech and the expected target [11, 12]. In this respect, an important contribution is an analysis based on Weinberger's Speech Accent Archive [13], which considers native and non-native varieties of English alike, to analyse how the ASR system *Otter.ai* performs in investigating the effect of syllable structures on the realisations of clusters and of vowel substitutions in relation to vowel spaces [14]. Another contribution to the analysis of learner speech based on ASR is an attempt at categorizing ASR errors in terms of phonological features ([15]) such as high or low vowels or coronal or labial consonants.

Large Language Models (LLMs) have been exploited to analyse learner errors but on written data. For speech data, audio LLMs have been used to analyse accented speech recognition [16]. Previous research on Whisper [?] related to accentedness includes a comparison of ASR performance (including Whisper) for Indian native and non-native speech [17]. Experiments in audio LLMs with accented speech has mostly focused on customising text-to-speech systems [18, 19]. For example, VALL-E X [20] produces "cross-lingual speech synthesis", a synthetic text-to-speech in a foreign language with a textual prompt and a sound sample of a speaker's voice. Previous research on L2 speech has investigated the potential uses of ASR for diagnoses but not in an automatic way. Current Second Language Acquisition on Automatic Speech Recognition (ASR) concur on the importance of phone substitution [11, 12] for the main L2 detected errors. [21] have also used wav2vec [10] and Whisper with k-means but with MFCC representations. To the best of our knowledge, our paper is the first

---

[1] https://alphacephei.com/vosk/

paper that uses Whisper probability scoring to investigate learner speech. Using a C++ implementation of Whisper [22], hereafter referred to as 'Whisper', we explore Whisper's internal representations with the probabilities associated by the system to the transcription task and to the language detection task detailed in [1]. Contrary to previous research on automatic phone-level pronunciation scoring (including the 'Goodness of Pronunciation' (GOP) [23] measure) this type of language scoring is not measured in relation to human judgements with forced alignment.

Whisper is a multilingual speech, large language model that has been trained for transcription and translation with thousands of hours of speech. It also has several multilingual models, as well as .en models trained with only English data designed to transcribe English exclusively. Table 1 lists the models used in this study.

Size	Parameters	Required VRAM	speed
tiny	39 M	1 GB	32x
base	74 M	1 GB	16x
small	244 M	2 GB	6x
medium	769 M	5 GB	2x
large	1550 M	10 GB	1x
large-v1	1550 M	10 GB	1x
large-v2	1550 M	10 GB	1x

Table 1: Whisper models tested for our experiments

Three main differences distinguish Whisper from ASR. First, Whisper has several models and may be able to produce both fine-grained interpretations of the speech signal with a `tiny` model and a potential gold standard with a `medium` model. The second difference is related to the multilingual training and the resulting language detection feature. Finally, within the C++ experimental implementation [2], some specific parameters on which the final output is based can be accessed.

We explore how two Whisper functionalities (assigning a probability to the subtoken, language detection) can be used for non-native scoring and indirect analysis of pronunciation errors. The rest of the paper is structured as follows. Section 2 presents our method and the data we used. Section 3 presents our results and Section 4 discusses them. Section 5 concludes.

2 Experimental design

For our experiments, we first present the linguistic data we used: read speech collected in a university to compare the Whisper transcriptions by the different models to the reference text (see Appendix A). We also used two published learner corpora that had reference points for levels, the ISLE corpus [4] and the ANGLISH corpus [2] to test the Whisper functionalities to classify learners. We then present the main Whisper features we tested and the two metrics that we have used, the classic word error rate and the Levenshtein distance.

---

[2] https://github.com/ggerganov/whisper.cpp

2.1 Testing data

38 French learners of English (read speech) For our first experiment, 25 female speakers and 13 male French speakers from the University of Limoges read two paragraphs from Conrad's *Typhoon* (see Appendix A)[3]. The Whisper .txt file outputs for the transcription and translation tasks were compared to the original text that was read by these second-year undergraduates. We used WER (Word Error Rate) and Levenshtein distance to assess the various graphic representations of the reference text as produced by Whisper (the .txt output of the LLMs) and to show how the differences can be transformed into operationalised feedback for learners. We will suggest that many spelling variations from the reference transcription can be turned into phonetic diagnoses for segmental or suprasegmental errors.

60 speakers from the ANGLISH corpus (spontaneous speech) Because we needed data tested for the levels of the learners, we also resorted to another corpus, the ANGLISH corpus, designed by Anne Tortel [2]. She selected three types of speakers (native speakers, advanced and intermediate speakers of English) to test the scoring. The ANGLISH corpus is a corpus designed by Anne Tortel as part of her PhD and it corresponds to 60 monologues, by 10 men and 10 women who were recorded in an an-echoic chamber to represent three levels of proficiency. FR1 corresponds to an intermediate level of English; these are subjects who stopped using English after the baccalaureate. The other group called 'FR2' is meant to represent advanced learners of English; these were third-year undergraduate students at the University of Aix-en-Provence. The GB group corresponds to recordings of 20 anglophones, who had either a British or an American pronunciation. We used the Whisper `tiny` model and `large` models to produce their scores and our experiments consist in trying to see if the average of the confidence scores produced by Whisper for each of the subtokens can be used to predict learner levels with the assumption that we should be able to classify three groups, FR1, FR2 and GB. We used the monologue task of the ANGLISH corpus, where subjects had to talk for two minutes.

26 German and Italian speakers from the ISLE corpus (read speech) We used the graded data from the Interactive Spoken Language Education (ISLE) corpus [4]. We re-organised the ELRA data compiled in 1999 in a unique dataset gathering metadata, prompts, objectives and expert annotations. We aggregated the sound files [4] of the different sections ("blocks") of the corpus per speaker. Three blocks correspond to the reading task of a passage from a novel and the other blocks aimed at evaluating more specifically phonemes, weak forms, consonant clusters, connected speech processes and word-class alternating pairs using isolated sentences. Three native experts graded the learners for the quality of their English, we report the biased distribution in Table 3.2 : the dataset of German speakers only included learners of level 3 and 4 and were more advanced than Italian speakers. We use this graded data to test the possibility of using Whisper to identify the native language or the level of the learner.

---

[3]We thank Maelle Amand for letting us use the dataset, described in [24].
[4]The data was processed from the ELRA distribution.

4

2.2 Whisper's customised C++ implementation

The various predictions are associated with a score and we can produce these for the different languages. Conversely, language predictions can be queried for multilingual models only. Whisper has also been trained with unique monolingual data in English, and these models can not be used to predict the language. With models trained with English data only (the .en models) and with the multilingual models, the C++ implementation allows the indirect display of probabilities. The colours correspond to the ten gradient values of probabilities, following the formula available in the code:

$$3^p * n\_colors$$

Figure 1 represents the realisations of a male learner as analysed by the `tiny` model. The final [BLANK . audio] transcription corresponds to a coda hallucination. Silent Final sequences (usually over 2s) may trigger hallucinations (thank you"). Note that tokens are subdivided in subtokens (byte pair encoding) and subtokens are not morphologically motivated, hence "*gir—lish*" for *girlish*. The red colour corresponds to uncertain transcriptions ("and his dish up" for "his knees up", see Appendix A). Our C++ implementation allows the exportation of these subtokens and their probabilities. We use another specific feature of large language models that distinguishes them from automatic speech: the fact that the linguistic data is encoded into subtokens in a phenomenon known as 'byte-pair encoding' (BPE). These byte-pair encoded units or 'subtokens', as we call them in the rest of the paper, are associated with probability prediction scores. The colour visualisation could be used in computer-assisted pronunciation teaching (CAPT) systems to provide learner feedback on their phonetic realisations. The probabilities correspond to the probability that the subtoken predicted by the LLM is true, and is presented by Gerganov as "confidence" of the model [22]. Figure 1 represents the realisations of a male learner as analysed by the `tiny` model. The final [BLANK . audio] transcription corresponds to a coda hallucination. Silent Final sequences (usually over 2s) may trigger hallucinations ("thank you"). Note that tokens are subdivided into subtokens (byte pair encoding) and subtokens are not morphologically motivated, hence "*gir—lish*" for *girlish*. The red colour corresponds to uncertain transcriptions ("and his dish up" for "his knees up", see Appendix A). Our C++ implementation allows the exportation of these subtokens and their probabilities. We use another specific feature of large language models that distinguishes them from automatic speech: the fact that the linguistic data is encoded into subtokens in a phenomenon known as 'byte-pair encoding' (BPE). These byte-pair encoded units or 'subtokens', as we call them in the rest of the paper, are associated with probability prediction scores. The colour visualisation could be used in computer-assisted pronunciation teaching (CAPT) systems to provide learner feedback on their phonetic realisations. The probabilities correspond to the probability that the subtoken predicted by the LLM is true, and is presented by Gerganov as "confidence" of the model [22].

Fig. 1: Whisper confidence estimation of transcription subtokens `tiny` model, C++ implementation [22]

2.3 Evaluation metrics

To evaluate the transcriptions produced by the Whisper models, we used the standard metric for ASR (Automatic Speech Recognition), WER (Word Error Rate) which can be defined as (1) [25]

$$WER = \frac{S + D + I}{N}. (1)$$

where:

- S represents the number of substitutions (errors where a word is replaced),
- D represents the number of deletions (errors where a word is missing in the hypothesis but present in the reference),
- I represents the number of insertions (errors where an extra word is present in the hypothesis but not in the reference),
- N represents the total number of words in the reference.

6

The Levenshtein distance [26] between two strings a and b can be defined recursively as follows:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 & \text{otherwise} \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} \end{cases} \quad (1)$$

where

- $\text{lev}_{a,b}(i,j)$ is the Levenshtein distance between the first i characters of a and the first j characters of b.
- a_i and b_j are the characters at positions i and j in strings a and b, respectively.
- $1_{(a_i \neq b_j)}$ is an indicator function that returns 1 if a_i is different from b_j, and 0 otherwise.

One of the contributions of our paper is that we tried to asses the distance between a reference text and a transcription on a global level, but also at a token level, using a specifically designed alignment algorithm.

2.4 A greedy alignment algorithm

Because large language models require a subtokenization process sometimes called 'byte pair encoding' (BPE), we needed to create a script that would re-align not only the Whisper ASR transcriptions but also the subtokens produced by the LLM. To address the evaluation of the quality of the English on read speech where reference labels are available, we wrote an R [27] script to compare the Whisper predictions at the subtoken level with the reference text and compute the Levenshtein distance at the token level. Figure 2 presents an output of this algorithm. The first column corresponds to the reference text token, the second column corresponds to the textual output of the tokenized as a column. The pipe separates the different subtokens (as in *Obs—erving*). The third column corresponds to the merger of the different subtokens of the predictions, and the id blocks correspond to the strings of subtokens that do not match the reference transcription. This algorithm is a potential backbone for the detailed investigation. The probability of the subtoken and the time stamps can be extracted, so that further analyses could investigate the signal with the time stamps corresponding to putative transcription errors.

We have realigned the predictions on a token level, so that we have been able to reconstruct the tokens out of the subtokens. This corresponds to the reference text. The second column corresponds to the subtokenized text and the corresponding probabilities that are assigned to this. Then we established the gap text, namely, the mismatches between the reference and the subtokenized text. We also report a string distance measure (the Levenshtein distance) between the reference token and the corresponding subtokens produced by Whisper. The algorithm is a greedy algorithm that tries to match the Whisper predictions expressed in subtokens with the reference text using punctuation symbols as reference points. As the discussion will show, this has consequences when a discrepancy between the number of commas can

7

(a) Source alignment

(b) Mistranscription alignment

Fig. 2: Comparison of the two types of outputs produced by the alignment algorithm. The realisations of the reference text tokens can be queried with the alignment of the Whisper transcriptions to the reference text (source alignment, 2a). The transcription mismatches can be investigated with the second alignment (Mistranscription alignment, 2b) that captures the mistranscriptions, enabling the computation of the Levenshtein distance

be observed in the transcriptions, but we managed to retain the initial texts with an align role function that corresponds to the tokenization of the reference texts. With the raw alignment, we can query the realizations of the different reference tokens by looking at the transcriptions of the different models using the same token in queries. Since it entails a situation with one-to-many and many-to-one, the align filter() makes it possible to visualize the situations of many-to-one. This may correspond to some phonological phenomena such as "pale and" being reanalyzed as a clitic form "palen", and it can also be another mismatch in terms of punctuation symbols, reinterpreted as a comma, a semicolon like before "another" being retranscribed by Whisper as a full stop, since in our data we do not find any semicolons in the Whisper transcriptions. Two functionalities are implemented in this alignment algorithm. One allows for a strict comparison of the tokens from the point of view of the reference text and can be used to produce queries to analyze difficult words for learners that are probably mispronounced in the data. The second one might be more interesting for the analysis of the performance of the different models, since it focuses on the discrepancies between the transcriptions produced by the different models, the reference alignment focusing on the segments that do not match between the Whisper transcription and the reference text. The column for refsize in the aligned version tells us how many linguistic tokens there are in the reference text column, and then API size indicates the number of tokens in the Whisper transcribed text. In this dataset, refsize is always 1.

The "Source alignment" allows the querying of the realizations of the individual tokens of the source text, whereas the "Mistranscription alignment" is target-oriented and looks at the mismatches observed with the initial reference text and the potential reanalysis, either phonological or lexical. In the discussion section, we insist on the fact that (the greedy algorithm being based on punctuation), if a punctuation mark is missing, it will look ahead in the transcription to find it. The detailed accuracy analysis of this prototype algorithm is beyond the remit of this paper, but we can predict that misalignments are more likely to be found when there are many punctuation signs in the reference text.

8

3 Results

3.1 Experiment 1: capturing learner data realisations with Whisper's 12 models

Our results for our first experiment in Table 2 confirmed the relevance of the `medium` model as being closest to the reference text if we take the Levenshtein distance (LD) as reference. If we take into account WER, we observe that the difference between the `tiny` and `tiny.en` models is only marginally significant (t-test, p=0.03) for WER but the difference between the `tiny` and the `large` model is very significant ($p < 0.001$) for WER. For Levenshtein distance, the difference is significant between the `tiny` and the `tiny.en` models ($p < 0.01$).

Table 2: Average distances to the reference text read by the learners according to the multilingual or English (.en) models whisper expressed as Word Error Rate (WER), standard error of the Word Error Rate (SE), number of substitutions (sub), insertions (ins), deletions (del), number of tokens (nbtok) or Levenshtein distance (LD)

model	WER	(SE)	sub	ins	del.	nb.tok.(ref)	nbtok (hyp)	LD
tiny	0.279	0.021	60	24	34	462	448	386
tiny.en	0.242	0.011	54	21	27	462	453	318
base	0.224	0.009	47	19	29	462	448	294
base.en	0.211	0.008	43	16	31	462	444	271
small	0.174	0.007	32	12	30	462	440	226
small.en	0.169	0.007	31	11	29	462	440	217
medium	0.140	0.006	24	9	26	462	442	189
medium.en	0.146	0.006	24	8	30	462	437	191
large	0.132	0.006	21	9	26	462	441	233
large-v2	0.132	0.006	21	9	26	462	441	233

This global analysis of the transcribed text by the different Whisper models suggests that the `tiny` model is the most likely to be informative about learner pronunciation, assuming deviations in the transcriptions correspond to deviations from the expected phonetic realisations. Because there was less training data for the model, it is less robust and probably more sensitive to speech variability. One of the possible explanations why larger models do not perform better than the `medium` model is that we did not use the normalisation script used in [1] to report WER results. Spelling variants like "sulphur"/"sulfur" "Herculian" / "Herculean" are not neutralised in our analysis.

3.2 The language detection feature

For this experiment, we exploit the two corpora that had different levels for the learners. For the ANGLISH corpus, the accuracy of the detection of the learner language is 80% for intermediate speakers FR1, but it should be borne in mind that the speakers were not tested for their phonetic proficiency during the data collection phase. Table reports the language predicted for the three levels. We explain why cy is predicted for

9

English speakers in the discussion section, but the crucial result is advanced speakers (FR2) are now predicted as having English speech in 90% of the cases.

Table 3: Distribution of the language predicted for the three groups of the ANGLISH corpus (monologue task)

predictions			
	cy	en	fr
FR1	0	4	16
FR2	0	18	2
GB	1	19	0

Table 4: Distribution of the language predicted by the Whisper large model for the ISLE data (aggregated sound file)

level	L1	large	count
1	ITALIAN	it	7
2	ITALIAN	en	6
2	ITALIAN	it	5
3	GERMAN	en	8
3	ITALIAN	en	4
4	GERMAN	en	15
4	ITALIAN	en	1

With the ISLE corpus, German learner speech files were predicted as being English speech input (see Table), but the corresponding learners also deemed to be of a higher level (3 or 4) by the ISLE annotators. Italian learner speech files were predicted as Latin (la), Slovenian (sl) or Italian by the `tiny` model. It should be noted that the size of the training data cannot explain why German learner productions were predicted as English speech, since German is more present in the training data (13,344 hours for multilingual speech recognition and 4309 for Translation). Latin was used only for the Translation task (1614 hours of training data), whereas Italian was used as training data for the multilingual speech recognition (2585 hours) and translation (2145 hours) as detailed in the appendix of [**?**]. For Italian learners' data of the lower level, the first language is detected with 100% accuracy. For Italian learners graded with level 2, a tie is observed between English and Italian. Further research is needed to estimate this observed threshold between the identification of the first language (here, Italian) and the identification of the target language (here, English). A systematic investigation of the phonetic realisations may account for these different predictions.

3.3 Exploiting subtoken probability

For the prediction of the language based on the levels of probability scores associated with subtokens, we first show that the means (Table 5) are consistent with the levels of the ANGLISH corpus. Each language level can be differentiated between the different levels by looking at the means and the standard error.

Table 5: Means and standard error per level in the ANGLISH data

group	mu	se
FR1	0.87	0.01
FR2	0.89	0.01
GB	0.94	0.00

Our hypothesis being that the sound files of higher levels are predicted with a higher probability, we also tested the reliability of these probability scores to assign speakers to a given level using the k-means algorithm, using an 80%-20% split of the data, 80% for training data, and 20% for testing data. For the ANGLISH corpus and its three levels, with k=3, the global accuracy is only 70% but the accuracy for the prediction of the GB is 90%. Higher probability scores reasonably correlate with proficiency. Table 6 shows the confusion matrix of the predictions (as rows) of the three levels of the ANGLISH corpus.

Table 6: Confusion matrix of the prediction of levels with the algorithm k-means with k=3 based on linguistic subtokens

	group		
pred	FR1	FR2	GB
FR1	13	6	2
FR2	5	11	0
GB	2	3	18

3.4 Evaluating realisations with the alignment algorithm

Searching for the transcription of individual tokens likely to be mispronounced shows how variable the graphic transcriptions of the Whisper models are. Previous research shows that only 16.1% of the mistranscriptions are common between the `tiny` and the `medium` models. Further research is needed to identify tendencies in the conflicting representations of the transcriptions produced by the Whisper models. Nevertheless, expected errors for learners can be observed in the transcriptions. *leaden* has a rarer

11

realisation of the <ea> digraph as /e/, as opposed to the most frequent realisations of the verb *lead*. This may explain forms like *leading* of *lid* (see Appendix C). Several reanalyses of *leaden* as *lead and* can be found. Clitic realisations of <and> are probable in the training data, and this probably also accounts for transcriptions such as *lid and*. Such segmental errors are easier to identify than suprasegmental errors [28] but some reanalyses can also be analysed as potential cues of suprasegmental errors. Many pentasyllabic realisations can be heard of *uncomfortable*, some of them with a tentative stress on the last but one syllable (see Appendix B), and stress placement may account for some the mistranscriptions. Another case of putative signalling of stress displacement is the reanalysis of polysyllabic words like *Herculean*, which tends to be transcribed with a determiner *a/her* because the first syllable is unstressed in the learner realisations. *Herculean*, which has a secondary stress /2010/, tends to be realised by learners as [0100], hence many reanalyses of the adjective as a determiner followed by a noun stressed on the first syllable like in *a Korean* (see other examples in Appendix D). Displaying these alternative forms ("her cutely and", "aircolion") could be a functionality incorporated in a dashboard for CAPT systems to signal potential stress placement errors.

4 Discussion

Our prototype alignment script needs to be more systematically tested, especially for reference texts with complex punctuation, as the algorithm relies on punctuation for the alignment and Whisper does not seem to produce semicolons in its transcriptions. More generally, the analysis of the predictions of multilingual audio LLMs requires a sophisticated way to account for the interactions between the speech input, the training data, the architecture of the model and the BPE subtokens. We report some of our observations on Whisper with learner data.

4.1 Duration ablation

We have not tested the effect of duration of the sound file on the values reported for the probabilities of the subtokens (larger contexts may entail higher probability scores). We have, however, noted that the language detection probability was sensitive to duration of the speech input. With the ISLE data, we split the aggregated file into two, three, four or five sections of equal duration. The probability reported for the beginning of each split was very similar, but the probability varied as evidenced in the boxplot (see Figure 3). Because the sound file aggregates the different prompts of the ISLE tasks, this variability can probably be explained by the fact that the learners varied in performance for each of the prompts.

Whisper's language detection feature should be used with caution, also because of the training data.

4.2 Potential biases in the training data

From our experience, some learners can be labelled as 'cy', i.e., categorised as Welsh speakers. This is because of an error in some of the labels of the training data (reported
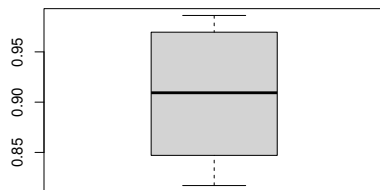
12

Fig. 3: Variability of the language detection probability reported when the duration of the ISLE sound file of learner #003 is divided into 2, 3 , 4 or 5

in [29]); some sound files labelled as 'cy' were actually recordings of English, so that Whisper has been partially trained to label speech as 'cy' for English data. The Whisper training data has not been made public, but some aspects of its bias can be observed from the outputs. The spelling is American and no semicolons were observed in our transcriptions. Using recordings of natives, the American speaker got a higher probability score of being English than a British speaker from the Midlands. More interestingly, Irish speakers from the county Galway with rising intonation in assertions also had lower probability scores in language detection for a prosodic pattern (Urban Northern British Intonation) less frequently found in American speech. It seems reasonable to assume that the training data is American-centric and that fine-tuning Whisper with different varieties of English could be useful before using Whisper for prosodic training in CAPT systems. Another expected bias of the training data is the use of read speech: some speakers do have a realisation of *Herculean* as a [0100] variant, but not in scripted speech, judging from examples on YouGlish [5] and this limited exposure to stress variability may account for the numerous reanalyses we observed.

One last potential source of bias is the byte pair encoding. We probed the subtoken dictionaries of the Whisper models, which are identical for all the models, but this does not imply that the mapping of the acoustic signal to subtokens is identical. The variability of the number of subtokens produced by each model suggests the opposite. The analysis of the individual accuracy of the probability score assigned to individual subtokens is a daunting task. First, polysyllabic infrequent words are likely to be represented by several subtokens. Second, we would need to test the adequacy of the model output in relation to phonetic variability. In other words, are audio LLMs acoustic models? As a starting point, we report the discrepancies of the transcriptions at the subtoken level in the next subsection where we compare three Whisper model outputs at subtoken level in relation to the speech signal.

---

[5] https://youglish.com/pronounce/Herculean/english?

13

4.3 Calibration curves and model response

The same speech input is transcribed differently by the Whisper models. Our results suggest that a correlation exists between a smaller Levensthein distance and a higher probability score of the subtokens. This does not imply that a high probability score is associated with a true prediction of the LLM. As an illustration of this point, we resorted to a method previously used to assess LLM predictions. For instance, [30] uses calibration curves to evaluate the accuracy of LLM statements regarding particular datasets and asserts that calibration serves as an additional measure for assessing the quality of the predictions. We focused on the best model to emulate the reference (the `medium` model has the smallest Levensthein distance) and the `tiny` and `tiny.en` models are the best candidates for the simulation of native misunderstandings. Previous results on WER discarding the normalisation scripts showed the `medium` model can do better on WER [31]. Table 2 shows the `medium` model does better for Levenshtein distance. To assess and visualise the quality of the model predictions, we manually annotated the accuracy of the subtoken predicted (positives, plotted on the Y axis), comparing the Whisper subtoken prediction to what can be construed from the sound file by a native speaker. The Y axis reports accuracy of the subtokens for `tiny` and `tiny.en` models, potential candidates for native simulated understanding of the speech input, and the `medium` model, a plausible candidate for the transcription of the target hypothesis (what the learner intended to say). Figure 4 plots the accuracy of the prediction according to the probability of the subtoken, which can be assimilated to the confidence that could be granted to the prediction. In calibration curves, the optimal model follows the x=y line (represented as a dotted line). The `tiny.en` model is closer to the ideal calibration x equals y (represented here as the dashed line), which suggests that it might be more sensitive to learner deviations from the expected realisations, assigning lower probabilities for transcriptions that are not correct. Conversely, the medium model, being more robust, produces subtokens that are mostly accurate. The most striking phenomenon is the "overpessimistic" pattern of the `medium` prediction for lower probabilities. Why would this model, being better, have some low probabilities in the predicted subtokens? Qualitatively analyzing the low probabilities of the medium predictions, we realised that many of them are actually subtokens forming part of linguistic tokens, so that there might be a form of architectural bias in the BPE where longer linguistic tokens are split into subtokens that are assigned lower probabilities.

5 Conclusion

For the data we tested, the native language of the lower-level learners can be predicted by larger Whisper larger models, but if English is predicted for a spoken input, the probability of such a prediction can not be used as a strict correlate of the learner level. The variability is too strong depending on duration of the speech input. In our data, a lower probability assigned to a subtoken is a potential plausible cue that the word was mispronounced. We show that this is consistent with the different learner levels, and that could be true for various L1s. We have shown the relevance of the scoring of the probability of the language course. This potential feedback for phonetic

14
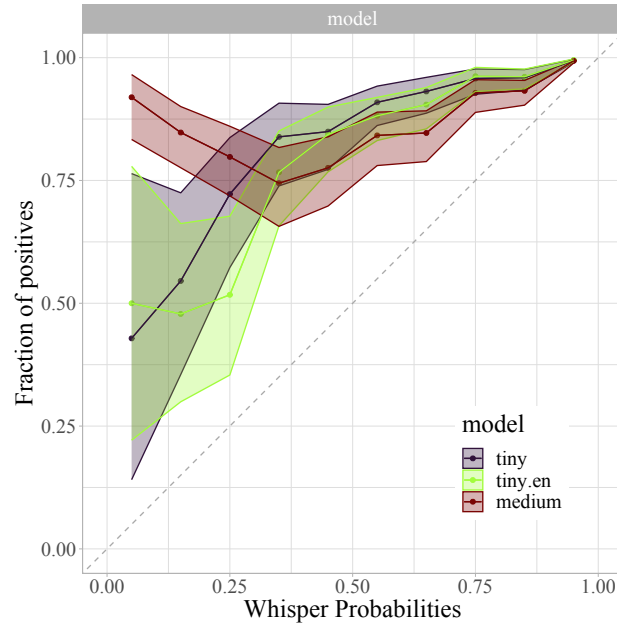
Fig. 4: Calibration curve for three Whisper models for the transcription of the learner #003 from the ISLE corpus

pronunciation errors based on the Whisper LLM is not L1-dependent, illustrating examples with French, German, and Italian learners of English. Some of the smallest Whisper models produce (mis)transcriptions that may be an adequate simulation of native misunderstanding of learner speech. This potential property should be systematically tested, for example comparing the recordings where "pigtail" was transcribed as "big tail" analysing VOT and pitch variability. More generally, the variability of the (mis)transcriptions across the different Whisper models of the same speech input needs to be explained for explainable artificial intelligence (XAI).

Supplementary information.

Declarations

- **Conflict of interest** The authors declare that they have no competing interests
- **Data availability** The ISLE corpus is available from ELRA https://catalogue. elra.info/en-us/repository/browse/ELRA-S0083/. The ANGLISH data is available on the public repository ORTOLANG https://www.ortolang.fr/market/corpora/ sldr000731/v2.
- **Code availability** The alignment script is available on https://github.com/ statsmaths/paper-replication.
- **Author contribution** Nicolas Ballier conceptualised the paper and wrote the first draft. Taylor Arnold wrote the alignment script and contributed to the ISLE data

15

Appendix A   Reference text of the reading task

Observing the steady fall of the barometer, Captain MacWhirr thought, "There's some dirty weather knocking about." This is precisely what he thought. He had had an experience of moderately dirty weather—the term dirty as applied to the weather implying only moderate discomfort to the seaman. Had he been informed by an indisputable authority that the end of the world was to be finally accomplished by a catastrophic disturbance of the atmosphere, he would have assimilated the information under the simple idea of dirty weather, and no other, because he had no experience of cataclysms, and belief does not necessarily imply comprehension. The wisdom of his county had pronounced by means of an Act of Parliament that before he could be considered as fit to take charge of a ship he should be able to answer certain simple questions on the subject of circular storms such as hurricanes, cyclones, typhoons; and apparently he had answered them, since he was now in command of the Nan-Shan in the China seas during the season of typhoons. But if he had answered he remembered nothing of it. He was, however, conscious of being made uncomfortable by the clammy heat. He came out on the bridge, and found no relief to this oppression. The air seemed thick. He gasped like a fish, and began to believe himself greatly out of sorts.

The Nan-Shan was ploughing a vanishing furrow upon the circle of the sea that had the surface and the shimmer of an undulating piece of gray silk. The sun, pale and without rays, poured down leaden heat in a strangely indecisive light, and the Chinamen were lying prostrate about the decks. [...] Captain MacWhirr noticed two of them especially, stretched out on their backs below the bridge. As soon as they had closed their eyes they seemed dead. Three others, however, were quarrelling barbarously away forward; and one big fellow, half naked, with herculean shoulders, was hanging limply over a winch; another, sitting on the deck, his knees up and his head drooping sideways in a girlish attitude, was plaiting his pigtail with infinite languor depicted in his whole person and in the very movement of his fingers. The smoke struggled with difficulty out of the funnel, and instead of streaming away spread itself out like an infernal sort of cloud, smelling of sulphur and raining soot all over the decks.

Appendix B: Examples of Whisper mistranscriptions of *uncomfortable*

ref_text	model	gap_text	count
uncomfortable	medium_en	and comfortable	1
uncomfortable	medium_en	and comforted	1
uncomfortable	medium_en	comfortable	1
uncomfortable	medium_en	incompatible	1
uncomfortable	medium	"uncomfortable "	1
uncomfortable	medium	comfortable	1
uncomfortable	tiny.en	a comfortable	1
uncomfortable	tiny.en	and comfortable	4
uncomfortable	tiny.en	in campatable	1
uncomfortable	tiny.en	meant and capable	1
uncomfortable	tiny.en	of incontivable	1
uncomfortable	tiny.en	uncountable	1
uncomfortable	tiny.en	ungothable	1
uncomfortable	tiny	and comfortable	2
uncomfortable	tiny	comfortable	1
uncomfortable	tiny	incomparable	1
uncomfortable	tiny	incompatible	1
uncomfortable	tiny	main and comfortable	1
uncomfortable	tiny	meant and comfortable	1
uncomfortable	tiny	uncorruptable	1
uncomfortable	tiny	ungovernable	1

Appendix C : Examples of Whisper mistranscriptions of *leaden*

ref_text	model	gap_text	count
leaden	medium_en	, leading it	1
leaden	medium_en	laden	1
leaden	medium_en	lead and	2
leaden	medium_en	leading	4
leaden	medium_en	leading it	1
leaden	medium_en	linen and	1
leaden	medium_en	unleaded	1
leaden	medium	, leading	2
leaden	medium	, leading heads	1
leaden	medium	, leading it	1
leaden	medium	, let on	1
leaden	medium	, letting it	1
leaden	medium	-laden heats	1
leaden	medium	a laden	1
leaden	medium	laden	4
leaden	medium	laden , ate	1
leaden	medium	laden heats	2
leaden	medium	laden hits	1
leaden	medium	lead and	1
leaden	medium	lid and	3
leaden	medium	lid and hip	1
leaden	medium	lid and hit	1
leaden	tiny.en	, leading	1
leaden	tiny.en	-leading	1
leaden	tiny.en	a little	1
leaden	tiny.en	laddened	1
leaden	tiny.en	lead and	15
leaden	tiny.en	lead and hits	1
leaden	tiny.en	lead in	3
leaden	tiny.en	leading	2
leaden	tiny.en	leading heads	1
leaden	tiny.en	leading it	1
leaden	tiny.en	lid and hit	1
leaden	tiny.en	little	1
leaden	tiny	, leading	2
leaden	tiny	, leading heads	1
leaden	tiny	, leading hits	1
leaden	tiny	, leading it	1
leaden	tiny	, led	1
leaden	tiny	, lit	1
leaden	tiny	foredown led	1
leaden	tiny	laden	3
leaden	tiny	laden hits	1
leaden	tiny	lead and	11
leaden	tiny	lead in	2
leaden	tiny	leading hits	1
leaden	tiny	linen , hit	1
leaden	tiny	on lead and	1
leaden	tiny	powered and ledon	1
leaden	tiny	the hidden hits	1
leaden	tiny	the lead in	1
leaden	tiny	the unleading hits	1

18

Appendix D : Examples of Whisper mistranscriptions of *Herculean*

ref_text	model	gap_text	count
herculean	medium_en	a Korean	1
herculean	medium_en	a cholera	1
herculean	medium_en	a kulean	1
herculean	medium_en	achilles	1
herculean	medium_en	arkadian	1
herculean	medium_en	her Korean	1
herculean	medium_en	her chilean	1
herculean	medium_en	her clean	3
herculean	medium_en	her culean	1
herculean	medium_en	her curling	1
herculean	medium_en	hickory	1
herculean	medium	Herculean	1
herculean	medium	a Cullian	1
herculean	medium	aculure on	1
herculean	medium	arcane	1
herculean	medium	arculean	1
herculean	medium	curly	1
herculean	medium	her Acheulean	1
herculean	medium	her clean	3
herculean	medium	her curly on	1
herculean	medium	her killian	1
herculean	medium	heroclone	1
herculean	medium	oculial	1
herculean	tiny.en	Arkalian	1
herculean	tiny.en	a crayon	1
herculean	tiny.en	a curian	1
herculean	tiny.en	a curling	1
herculean	tiny.en	a hate -culline	1
herculean	tiny.en	aculean	1
herculean	tiny.en	aculian	1
herculean	tiny.en	air -culion	1
herculean	tiny.en	ekulean	1
herculean	tiny.en	her cally and	1
herculean	tiny.en	her clean	2
herculean	tiny.en	her curling	1
herculean	tiny.en	her curly and	1
herculean	tiny.en	her gluing	1
herculean	tiny.en	her kiln	1
herculean	tiny.en	herculine	1
herculean	tiny.en	oculi and	1
herculean	tiny.en	our culian	1
herculean	tiny.en	the hakiran ,	1
herculean	tiny	a chulian	1
herculean	tiny	a chulian	1
herculean	tiny	a clean	1
herculean	tiny	a crayon	1
herculean	tiny	a heckling	1
herculean	tiny	aircolion	1
herculean	tiny	echelion	1
herculean	tiny	her Qulian	1
herculean	tiny	her chulian	2
herculean	tiny	her chuling	1
herculean	tiny	her clean	2
herculean	tiny	her collier and	1
herculean	tiny	her culean	1
herculean	tiny	her culey and	1
herculean	tiny	her curly and	2
herculean	tiny	her cutely and	1
herculean	tiny	her killer and	1
herculean	tiny	her killing	1
herculean	tiny	herculey and	1
herculean	tiny	herculian	1
herculean	tiny	herkali	1
herculean	tiny	herkilling	1
herculean	tiny	herkilly and	1
herculean	tiny	herkul and	1
herculean	tiny	hurtly enchilers	1
herculean	tiny	our Qliian	1
herculean	tiny	percolian	1
herculean	tiny	to be covered	1
herculean	tiny	were her killion	1

References

[1] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[2] Tortel, A., Hirst, D.: Rhythm metrics and the production of English L1/L2. In: Proc. Speech Prosody 2010, p. 959 (2010). https://doi.org/10.21437/SpeechProsody.2010-49

[3] Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., Souter, C.: The ISLE corpus of non-native spoken english. In: Proceedings of LREC 2000: Language Resources and Evaluation Conference, Vol. 2, pp. 957–964 (2000). European Language Resources Association

[4] Atwell, E., Howarth, P., Souter, D.: The isle corpus: Italian and german spoken learner's english. ICAME Journal: International Computer Archive of Modern and Medieval English Journal **27**, 5–18 (2003)

[5] Rogers, C.L., Dalby, J.M., DeVane, G.: Intelligibility training for foreign-accented speech: A preliminary study. JASA **96**(5), 3348 (1994) https://doi.org/10.1121/1.410623

[6] Watson, C.S., Reed, D.J., Kewley-Port, D., Maki, D.: The indiana speech training aid (istra) i: Comparisons between human and computer-based evaluation of speech quality. Journal of Speech, Language, and Hearing Research **32**(2), 245–251 (1989)

[7] Dalby, J., Kewley-Port, D.: Explicit pronunciation training using automatic speech recognition technology. CALICO **16**(3), 425–445 (1999)

[8] Inceoglu, S., Chen, W.-H., Lim, H.: Assessment of L2 intelligibility: Comparing l1 listeners and automatic speech recognition. ReCALL **35**(1), 89–104 (2023) https://doi.org/10.1017/S0958344022000192

[9] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., *et al.*: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011). IEEE Signal Processing Society. https://infoscience.epfl.ch/record/192584/files/Povey_ASRU2011_2011.pdf

[10] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[11] Chanethom, V., Henderson, A.: Alignment in ASR and L1 listeners' recognition of L2 learner speech: A replication study. In: 15th International Conference on

21

Native and Non-native Accents of English, Łódź, Poland (2022). Université de Łódź. https://hal.science/hal-03929160

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

[13] Weinberger, S.: Speech accent archive. George Mason University. http://accent.gmu.edu (2015)

[14] Chan, M.P.Y., Choe, J., Li, A., Chen, Y., Gao, X., Holliday, N.: Training and typological bias in ASR performance for world Englishes. In: Proc. Interspeech 2022, pp. 1273–1277 (2022). https://doi.org/10.21437/Interspeech.2022-10869

[15] Arora, V., Lahiri, A., Reetz, H.: 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 (2018)

[16] Aksënova, A., Chen, Z., Chiu, C.-C., Esch, D., Golik, P., Han, W., King, L., Ramabhadran, B., Rosenberg, A., Schwartz, S., et al.: Accented speech recognition: Benchmarking, pre-training, and diverse data. arXiv preprint arXiv:2205.08014 (2022)

[17] Javed, T., Joshi, S., Nagarajan, V., Sundaresan, S., Nawale, J., Raman, A., Bhogale, K., Kumar, P., Khapra, M.M.: Svarah: Evaluating English ASR Systems on Indian Accents. In: Proc. INTERSPEECH 2023, pp. 5087–5091 (2023). https://doi.org/10.21437/Interspeech.2023-2588

[18] Jiang, Z., Ren, Y., Ye, Z., Liu, J., Zhang, C., Yang, Q., Ji, S., Huang, R., Wang, C., Yin, X., Ma, Z., Zhao, Z.: Mega-TTS: Zero-Shot Text-to-Speech at Scale with Intrinsic Inductive Bias (2023)

[19] Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A.: Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In: International Conference on Machine Learning, pp. 2709–2720 (2022). PMLR

[20] Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling (2023)

[21] Islam, E., Park, C., Hain, T.: Exploring speech representations for proficiency assessment in language learning. In: 9th Workshop on Speech and Language Technology in Education (SLaTE) Proceedings, pp. 151–155 (2023). International Speech Communication Association (ISCA)

[22] Gerganov, G.: whisper.cpp : A high-performance inference of OpenAI's whisper

automatic speech recognition (ASR) model (2003)

[23] Witt, S.M., Young, S.J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication **30**(2-3), 95–108 (2000)

[24] Ballier, N., Méli, A., Amand, M., Yunès, J.-B.: 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), pp. 282–292 (2023)

[25] Martin, A., Daniel, E., Ward, N.: The use of the word error rate for evaluating automatic speech recognition systems. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing **1**, 77–80 (1998)

[26] Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**, 707–710 (1966)

[27] Team, R.C.: R: A language and environment for statistical computing [computer software manual]. Vienna, Austria (2024)

[28] Ballier, N., Martin, P.: Speech annotation of learner corpora. In: Granger, S., Gilquin, G., Meunier, F. (eds.) The Cambridge Handbook of Learner Corpus Research, pp. 107–134. Cambridge University Press, Cambridge (2015)

[29] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[30] Levinstein, B.A., Herrmann, D.A.: Still no lie detector for language models: Probing empirical and conceptual roadblocks. Philosophical Studies, 1–27 (2024)

[31] Ballier, N., Namdarzadeh, B., Zimina-Poirot, M., Yunès, J.-B.: Translating Dislocations or Parentheticals: Investigating the Role of Prosodic Boundaries for Spoken Language Translation from French into English. In: Machine Translation Summit 2023, vol. 19, pp. 119–131 (2023)

23