Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English

Nicolas Ballier, Adrien Méli, Maelle Amand, Jean-Baptiste Yunès

Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English

Nicolas Ballier[1 2], Adrien Méli [2], Maelle Amand [3], Jean-Baptiste Yunès [4]
[1]LLF / [2]CLILLAC-ARP / [4]IRIF Université Paris Cité, F-75013 Paris, France
[3] CeRES / Université de Limoges 39E rue Camille Guérin F-87000 Limoges
{nicolas.ballier, jean-baptiste.yunes}@u-paris.fr, adrienmeli@gmail.com
maelle.amand@unilim.fr

Abstract

This paper reports on a pilot study to use Whisper's large language model (LLM) as a tool for potential representation of segmental (phone) pronunciation errors. We compared the performance of the transcription outputs for the various models developed by the automatic speech recognition (ASR) system Whisper (Radford et al., 2022) ranging from 39 to 1,550 million parameters. We investigated 38 recordings of two paragraphs from Conrad's *Typhoon*. The whisper transcriptions were compared to the original text that was read by these second-year French undergraduates. We used WER (Word Error Rate) and Levenshtein distance to assess the various graphic representations of Conrad's reference text. We show how the differences can be transformed into operationalised feedback for learners. We used expert phonetic knowledge to check the plausibility of the phonetic interpretation with the signal (in particular the recall of H dropping produced by French learners). Our findings suggest that the transcriptions produced by the `medium` model converge with what a native speaker understands and that the `tiny` model produces alternate transcriptions that are plausible candidates for learner errors.

1 Introduction

Whisper is an audio multilingual large language model (mLLM) which can be used for two types of tasks, transcription (speech to text) and translation (only to English). Using thousands of hours of training data, mostly from Librispeech (Panayotov et al., 2015), a dataset of read speech of public domain books, Whisper has been trained with both multilingual data and English only data. Several models have been created with an increasing number of parameters, as listed in Table 1 (Radford et al., 2022). Probably because of Named Entity Recognition (NER) issues as acknowledged in (Radford et al., 2022), proper nouns (but other tokens as well) can undergo what we call a *retranscription*, i.e., that differs from the original text but that is phonetically consistent with the speech input, e.g., *Macquaire* instead of the expected *McWhirr*.

In this paper, we follow the standard phonological convention that indicates graphemes (letters) with angled brackets(<>), realisations in square brackets and phonemes (or targets) with slanted bars (//). Our research questions are as follows: do ASR retranscriptions differ from one Whisper model to the next, and how realistic are they as (re)interpretation of learner phonetic realisations?

Previous research has suggested that the Whisper retranscriptions vary across Whisper models (**?**) while trying to be faithful to the phonetic input of a foreign pronunciation. This paper essentially assesses two Whisper models (`tiny` and `medium`) in their ability to capture relevant L2 pronunciation errors in classroom or computer-assisted learning environments. We want to test the hypothesis that the `tiny` model is more likely to retranscribe pronunciation errors than the `medium` model. Our hypothesis is somehow counter-intuitive as the lowest model with the least number of parameters is chosen to be the most efficient to represent / to emulate learner representation or the learner data as perceived by native speakers. We are working on the discrepancy between the transcriptions integrated condition with the reference target hypothesis.

We first provide a quantitative analysis of these discrepancies before analysing the fine phonetic renditions of the different files. Two professionals trained in phonetics analysed the phonetic data and tried to extract one of the most striking features from a phonetic point of view in order to be used as feedback for learners : H-dropping, namely, the lack of aspiration. Since /h/ is not part of the phonemic inventory of French, most learners either omit the sound or substitute it with a glottalisation (Exare, 2022). The two operations were carried out independently. We then analyse the extent in which

Whisper's graphic renditions match the phonetic interpretation of the learners' mispronunciations.

The rest of the paper is structured as follows: Section 2 presents the previous research carried out on automatic speech recognition with learner data. Section 3 presents the data we tested and the metrics we used. Section 4 presents our results and Section 5 discusses them.

2 Previous Research

The use of ASR in pronunciation training dates back to the 1990s. A preliminary study pioneered the use of ASR in L2 pronunciation (Rogers et al., 1994), showing that ASR helped improve intelligibility in the learner's L2 and that the improved targeted phonetic contrasts (/iː/ vs. /ɪ/, /θ/ vs. /s/) were also found in untrained words. Watson et al. (1989) compared human and ASR evaluations of speech quality. Some explored ways to integrate ASR in pronunciation training programs (Dalby and Kewley-Port, 1999), while others focused on the creation of feedback derived from the ASR transcriptions. More recent studies (Inceoglu and Lim, 2023) 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 (Povey et al., 2011), Vosk [1], wav2vec 2.0 (Baevski et al., 2020), and others for ASR models.

ASR models have also been applied to the analysis of L2 speech. Previous studies focused on the discrepancies between the ASR output of L2 speech and the expected target (Chanethom and Henderson, 2022; Inceoglu et al., 2020). In this respect, an important contribution is an analysis based on Weinberger's Speech Accent Archive (Weinberger, 2015), 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 (**?**).

To the best of our knowledge, our paper is the first paper that uses Whisper to investigate learner speech and, more generally, that compares the performance of several models within the same ASR system.

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

Size	Parameters
tiny	39 M
base	74 M
small	244 M
medium	769 M
large	1550 M
large-v2	1550 M

Table 1: Whisper's main models for speech recognition, after (Radford et al., 2022)

3 Materials and Method

3.1 Whisper Parameters and Outputs

Whisper uses the Encoder-Decoder Transformer architecture and takes audio as input, chunked in 30s windows and converted to a log-Mel spectrogram. Whisper is trained to predict the corresponding text (Radford et al., 2022) and its translation into English. Transcription and translation are the two main tasks, but Whisper can also provide language identification. We tested the learner speech with Whisper's different models. Table 1 lists the corresponding parameters of these models. After a transcription, each Whisper model outputs files in the Hugging Face implementation with with or without time stamps. A .json file includes the metadata of the prediction outputs for each segment (the average log probability, the compression ratio and the probability of the absence of speech).

3.2 Selected Reference Target for Learner Data

38 graduate-level learners of English at a French University were asked to read the first two paragraphs of chapter 2 from Joseph Conrad's *Typhoon* (1902).[2]. The text counted 408 words with 17 sentences. It was deemed suitable for L2 speakers with a C1 level by CEFR standards by the CATHOVEN text analyser [3] due to the richness of the vocabulary and complexity of the sentences. The high cognitive load required to read the text was expected to highlight pronunciation difficulties that are not fully mastered by the L2 learners (Christodoulides, 2016). These two paragraphs contain a wide array of potential pronunciation difficulties for French

[2] The students were warned that the term *Chinaman* was considered offensive and that it should not be used today when referring to a person.

[3] https://hub.cathoven.com/?scene=analyser

L2 learners (voicing of intervocalic <s> in *precisely*, H-dropping of initial and medial /h/ or H-intrusion (*hair* for *air)*, unstable vowel length contrast (*(h)it* instead of *heat*), lack of initial aspiration for voiceless plosives (*pigtail* is understood as *big tail*, vowel reduction, misplacement of lexical stress...).

3.3 Metrics

To analyse the retranscriptions produced by Whisper, we used word error rate (WER) a standard metric for Automatic Speech Recognition systems and Levenshtein distance (Levenshtein et al., 1966), as produced by the R package {phonics}(Howard II, 2020), since it offers insights into the discrepancy between the target hypothesis and the learner realisation, and a graphic rendition of the learner realisation produced by the different models.

4 Results

4.1 Selecting the Optimal Model for Leaner Data Transcription

In this section, we report our findings on the Whisper .txt outputs by ASR model. Figure 1 displays the boxplots corresponding to the WER of the different Whisper models. No significant difference in performance (WER) was found between the models specifically trained with English data (whether `tiny.en` or `medium.en`) and the multilingual models. A t-test revealed no significant difference between the multilingual `tiny` model and English-only `tiny.en` model (t = 2.1947, df = 37, p-value = 0.03454). While the WER between the multilingual `tiny` model and the `medium` model was deemed significant (t-test : t = 7.3121, df = 37, p-value < 0.001), that between the `medium` and the `medium.en` was not significant. Nevertheless, a more detailed comparison revealed that the `tiny` model produces a higher WER than the `tiny.en` model, wheareas the `medium.en` model had higher error rates than the `medium` model. This seems to suggest that the `tiny` model is the most efficient model in capturing non-native pronunciation oddities, while the equivalent model based on English only seems to normalise such oddities.

4.2 Number of Retranscriptions and Model Size

String distance was also examined between the models, and more specifically, the number of additions and the number of tokens that were outputted by each model but were not in the reference text. We found that the number of added tokens decreased almost linearly with the log of the number of parameters for each model from `tiny` to `medium` (Figure 2).

The different models produce different types of respelling (and in varying quantities). This is true for the `tiny` vs. `tiny.en` models but also for the `medium` vs. `tiny` models. We tried to test the separability of the tokens that were retranscribed by these models and used a Venn's diagram to categorise the different model reinterpretations of the same acoustic signal (Figure 4). The retranscriptions of the different models are not mutually exclusive, as the `medium` and the `tiny` models share 16.1 % of their retranscriptions, but they must not be understood as a simple numerical decrease of alternative respellings across models. In fact, they include different tokens that are not in the reference text. Further research is needed to investigate why the different Whisper models produce different graphemic representations, since the models are based on the same (sub)token dictionary after the Byte-pair encoding.

4.3 Plausibility of the Whisper Respelling

This subsection tentatively reports on the precision of the retranscription, by detailing the phonetic interpretation of respellings. The 38 `tiny` models produced 832 tokens differing from the original reference text, including one recording transcribed exclusively into French. Some hapaxes corresponded to mispronunciations such as <alphnicate> for *half-naked*,which are consistent with common features amongst non-native speakers: h-dropping (Exare, 2017), monophthonguisation of <a> in <naked> with harmonisation with the second vowel ([nikit]instead of /neɪkɪd/) the devoicing of final consonants (here, /t/ for /d/, cf. (Hutin et al., 2020)).

4.4 Precision and Recall

Assessing precision and recall of the phonetic error detection means answering the following questions : how many of the Whisper retranscriptions point to an actual pronunciation error (precision) and how many of the learners' pronunciation errors were captured in the Whisper transcriptions (recall)? In this paper, we do not address the precision and recall of the phonetic errors by the system, as it would require intensive manual phonetic annota-

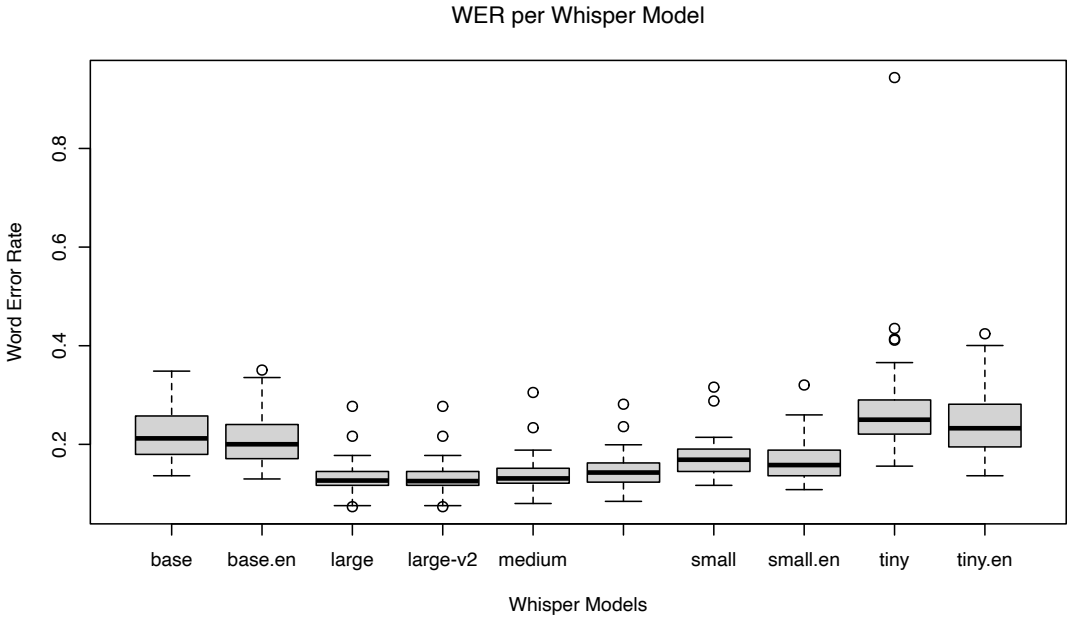

Figure 1: Dispersion of the WER across speakers for each Whisper model, trained on English data only (.en) or on multilingual data

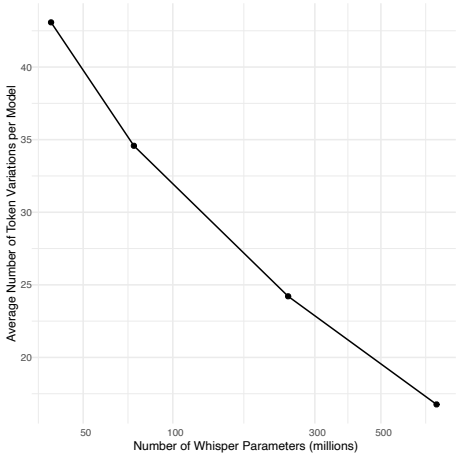

Figure 2: Relation between the numbers of parameters (log scale) of the `tiny`, `base`, `small` and `medium` models and the detection of Pronunciation Errors Candidates signalled by spelling variants or token additions

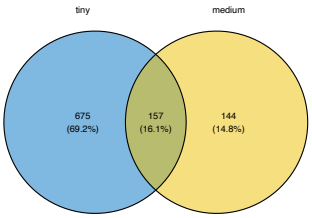

Figure 3: Venn's diagram of the common retranscriptions between the `medium` and the `tiny` models

tion of all the files, but we offer a critical diagnosis of some of the most frequent retranscriptions we observed, especially discussing those which potentially emulated some misunderstanding with a native. For precision, we analysed the 13 frequent added tokens and noticed false alarms for less frequent items as well. The presence of a reduced vowel for the realisation of *seaman* led to the transcription of the token as *semen* and as *seamen* (it should be noted that the Levenshtein distance is higher but that the two candidates for the learner realisation are homophonous). As can be seen in our inventory of most frequent retransciptions (Figure 4), some false positives can be observed: *grey/gray* for spelling divergences, *plowing/ploughing*, *sulphur/sulfur* and they account for half of the types of

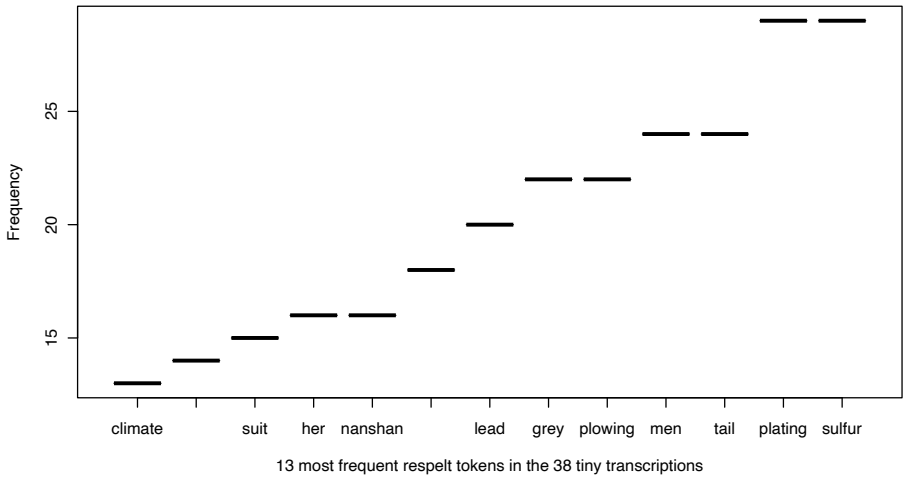

Figure 4: Top 13 aggregated retranscriptions of the `tiny` models

the 13 most frequent respelt tokens. Nevertheless, some frequent retranscriptions actually point to potential misunderstandings by native speakers of the realisations, as they correspond to fully-fledged minimal pairs revealing misrealisations such as *suit* for *soot*.

4.5 Recall of Non-native Phonetic Realisations by Whisper Retranscriptions : The Case of Aspiration

For our analysis of precision of recall, we focused on a very limited number of items in order to evaluate the plausibility of the Whisper retranscriptions, such as *languor* for *language*, and the retranscriptions of the aspiration of *heat* in the sequence *clammy heat*. As to the retranscription of *languor* as *language*, it is mostly due to the realisation as [gw] instead of [g] but the realisation of the final consonant arguably would not trigger misunderstanding for a native speaker. Two experts trained in English phonetics annotated the sound files for limited sequences in relation to our expectations about H-dropping (Exare, 2017). The two experts agreed with the Whisper transcriptions. For our analysis of recall, all the dropped /h/s in *clammy heat* were actually transcribed without an <h>. We auditorily investigated the unaspirated initial /p/ in *pigtail*, which were transcribed as *big*: they were pronounced without initial aspiration. More detailed acoustic analysis of Voice Onset Time (VOT) should be carried out to check the ability of the system to transcribe initial plosives in relation to expected values of Voice Onset Time in English (Abramson and Whalen, 2017; Lisker and Abramson, 1967) and French (Caramazza et al., 1973), in order to investigate whether threshold effects of VOT could be observed in relation to much lower VOT reference values for French (18ms for French /p/ against 59 ms for English).

5 Discussion

At this point, we are not quite able to characterise the different types of “sensitivity" of the Whisper models: tokens do not systematically trigger a spelling variation across all the models.

5.1 Reliability of the detection of error candidates

Some false positives were observed, based on spelling variants (*half-naked* is hyphenated by the Whisper outputs, not in the original). Extra hallucinated ASR errors were observed in the transcriptions in the context of false starts, repairs or repetitions, so that some tokens were repeated several times and occasional cases of coda hallucinations were noticed with *Thank you* or *hit the bell button* being transcribed instead of final silences. Our hypothesis for these cases of coda brittleness of the audio LLM is that part of the training data was initially online and if an end-of-signal cue is captured by the ASR, then this may me transcribed as what might have been left out in the training data.

5.2 Semantically Plausible or Phonetically plausible?

Sequences such as *clammy it* for *clammy heat* raise the question of the semantic plausibility in relation to *surprisal* (Mansfield, 2021), namely the probability of having a given token rather than another one. It seems that the speech inputs, i.e., the phonetic acoustic cues, have more importance in the next-token prediction than just the conditional probability, which would reflect on the semantic plausibility and this apparent dominance of phonetic plausibility seems to prevail over semantic plausibility. Further systematic research analysing surprisal needs to be undertaken, but for an initial estimate of how Whisper outputs may violate semantic plausibility for phonetic plausibility, we computed surprisal using the large language model BERT. We used this to check some of the outputs that were phonetically consistent with the input but semantically less likely: "*He was, however, conscious of being made uncomfortable by the clammy heat. He was, however, conscious of being made uncomfortable by the clammy it*." Even though its surprisal value is much higher with *it* (2.251) compared with the surprisal value for *heat* (0.003), faithfulness to the acoustic signal (absence of aspiration) was observed. This initial foray suggests Whisper outputs are potentially more consistent with phonetic input than with semantic input. In other words, we need to explore the *affordance* (Krunic et al., 2009) of the large language model to accommodate to the acoustic realisations of the learners. How much of the phonetic variability can actually be accommodated by the textual production?

5.3 Alternative Measures of Pronunciation Distance

We did not resort to more elaborated metrics and probably more cognitively grounded measures of pronunciation distance based on the Naive discriminant learning analysis suggested by (Wieling et al., 2014). We are sensitive to the arguments they put forward against Levensthein distance, especially the misalignments produced by the possibility of having reduced vowels. They explain that they have what they call “sensitive sound distances" for tokens like *Wednesday*, which can be realised in a certain number of ways, as two or three syllables. They exemplify the schwa reduction to show that the Levenshtein distance exaggerates the scores in relation to this type of phenomenon. We used the more classical Word Error Rate (WER), which was computed with R (Team, 2023) but we did not apply the normalisation procedure [4] which was used when reporting Whisper performances for WER in (Radford et al., 2022).

5.4 Retranscriptions or Plausible Scenarios for Misunderstanding?

As our examples show, some of the substitutions or respelling proposed by word substitutions and phone substitutions do not necessarily correspond to actual native misunderstandings. In this respect, there is an imbalance between monosyllabic words more likely to convey misunderstanding because of the number of potential minimal pairs (what is known as phonological neighbourhood density) than polysyllabic words, as our *languor* / *language* example seems to suggest. The system is probably biased towards detecting monosyllabic misrealisations more easily, but this also reflects a skewed distribution which can be observed in the language lab exercises, where monosyllabic minimal pairs are much more frequent than polysyllabic examples. Pre-trained generative models are trained to produce tokens, which explains why a word like *funnel* when pronounced initially by a learner as [fju:] becomes transcribed as *funeral*, as this is the closest approximation in spite of the extra syllable.

5.5 Further Validation Procedures

This section discusses potential validation procedures, other than perception tests on native speakers and more detailed acoustic analyses for the transcription of *heat* as *hit*. A list of anticipated phonetic/phonological transfers could potentially be used to serve as the rationale for a confusion matrix analysing the Whisper output and the ability of a graphemic representation to capture phonetic errors.

The ISLE corpus (Menzel et al., 2000; Atwell et al., 2003) has reference transcriptions and validation procedures, but for much shorter segments in carrier sentences such as "*I said wait, not bait*". This corpus of non-native speech also has a read passage by German and Italian speakers, but it has not been annotated by experts. Our preliminary tests with Whisper suggest that heavily-accented speakers are detected as speaking in another language than English and transcribed accordingly.

[4] https://pypi.org/project/whisper-normalizer/

More generally, Whisper has to be tested for other first language speakers, and maybe with other second languages, with the proviso that some languages have a much smaller training size.

5.6 XAI and the Knowledge of the LLM of Different Sizes

Our experiments with Whisper with other recordings suggest that the large-v2 works better, i.e., produces a transcription output which might be more accurate for more sophisticated words. What is the underlying "knowledge" captured in these representations? Is it probably because more data was taken into account in the training phase that "stigmatal and supra-stigmatal features" (`medium` transcription) get (accurately) transcribed as "segmental and supra-segmental features" in the `large-v2` transcriptions. How "linear" is the understanding of the largest models? Is the progression linear between the different transcriptions or can thresholds be observed?

6 Limitations and Further Research

The ASR transcriptions of mispronunciations seem more relevant for segmental features than for supra-segmental features, even though a certain form of chunking is actually captured by a mix of punctuation symbols such as comma and full stops. This means that the word *, however,* in isolation can actually be analysed in terms of successful chunking. Part of the phrasing can be captured by the system through punctuation and, moreover, probably in an even more complex manner, as the end-of-the-line character also of a Whisper transcription corresponds to a form of prosodic chunking different from what is transcribed by a comma or a full stop. In any case, in terms of prosody, only tonality (the ability to properly chunk the prosodic units) can be analysed using Whisper. An important aspect of non-native realisations is the elusive ability to assign stress on the relevant syllables and, in that respect, only reanalyses can be used to track down stress misplacement, as is the case with *her Qulian* for *ˌHercuˈlean*, which is favoured by the stress misplacement. This ASR transcription reveals a weak vowel on *her*, making it more likely to be interpreted as the possessive pronoun. (Kamiyama and Amand, 2023) showed that a frequent incorrect lexical-stress placement amongst French L1 advanced learners of English is the placement of primary stress on the first syllable of words having a similar structure, such as *simulation*, *organisation*. However, unlike the students in Kamiyama & Amand (2023), the students of this study were enrolled in a pronunciation course with a strong focus on stress-imposing endings. The learner whose pronunciation led to the transcription of the form *her Qulian* may have treated the ending *-e.an* like the strong ending *-i.an*, which attracts lexical stress one syllable before the ending, i.e., *Braˈzi.li.an* (Kingdon, 1958).

6.1 Effect of the Training Data on the (Implicit) Rhotic Pronunciation Model

The Librispeech samples available on Hugging face[5] suggest a rather slow reading which is fully rhotic but possibly East coast of the United States (slight variation in the use of yod for *assumed*, *new* or *duke*). There may be a training bias and consequently an implicit rhotic pronunciation model with the data trained on Librispeech (Panayotov et al., 2015). As a baseline for native realisations, we tested the recording of the Librivox version read by Peter Dann, which exhibits a rhotic realisation [6]. The L2 learners of English in this study generally use both rhotic and non rhotic forms while reading the *excerpt from Typhoon*.

6.2 Gender Bias effects

Even though the system revealed that the performances were significantly different for male and female speakers, it is notable that the Levenshtein distances outputted by the `large` model and the `medium` model highlight diverging performances for male and female voices: the `large` model is slightly better than the `medium` for male speakers, but the `medium` model is noticeably better for female speakers.

6.3 Next Steps for ITSs

This subsection discusses how our findings could be implemented in Intelligent Tutoring Systems (ITS). Using an NVIDIA A100 GPU with 40 giga of RAMS, the transcription only took 5 minutes for all the models of two ISLE files, so that the Whisper system could be used to provide almost immediate feedback to learners (or post-hoc analysis when used in a virtual environment). Whisper tran-

[5] https://huggingface.co/datasets/librispeech_asr

[6] ttps://ia802507.us.archive.org/21/items/typhoonandotherstories_2206_librivox/

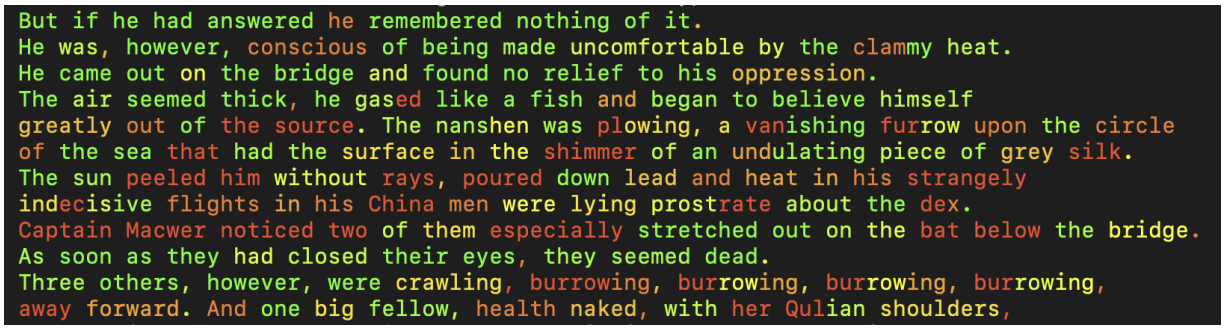

Figure 5: Confidence estimation of the predicted tokens as potential visual feedback from *Whisper.cpp* (Gerganov, 2003). Green: confident prediction, i.e., intelligible; red: least confident prediction, i.e., more phonetic training needed to be intelligible. Original text in appendix.

scriptions of non-native speech need to be tested on other tasks than read speech, even though the baseline can be established with the text that was read. With *unscripted*, i.e., spontaneous speech, we may use the `medium` transcription as baseline for the computation of the output of the `tiny` model. In that respect, the existence of several models is an important structural difference from other ASR systems such as *Otter.ai* whose current interface cannot produce a reference text to be compared with Otter's ASR output. For multi-speaker settings such as virtual environments or classroom interactions, speaker *diarisation* will have to be processed first, i.e., the creation of distinct transcribed segments when the speaker changes. Both *Otter.ai* and the experimental C++ implementation of Whisper provide speaker diarisation.

6.4 Scenarios for Potential Visual Feedback

Though experimental, the C++ implementation of Whisper called *Whisper.cpp* (Gerganov, 2003) allows fast processing of some of the Whisper parameters and a visualisation of the confidence estimation for the predicted tokens that is easily understood by teachers and students (Figure 5). The confidence scores are consistent with the phonetic realisations. Stress (mis)placement accounts for some of the scores, as *uncomfortable* was stressed on the penultimate syllable in this example. Running a recording of 151 seconds with its coloured transcription as output only took 4923.62ms on an M1 Pro processor. Feedback can be visually displayed shortly after the end of the recording. For further analyses, a more refined implementation could also output the corresponding confidence scores produced for each subtoken of the transcription (the coloured sequences correspond to the output of byte pair encoding and are not "words").

7 Conclusion

In this paper, we have shown that Whisper's LLM produces different outputs for the transcription task according to the different learner pronunciation models of a reference input. We showed that the number of parameters of the LLM models varied in relation to the detection of tokens varying from the reference text. A phonetic screening of part of the audio files showed the phonetic realism of the retranscriptions varying from the reference file (see appendix). For the analysis of L2 speech, the models trained with fewer parameters paradoxically do a better job at pinpointing L2 pronunciation misrealisations, as they seem more sensitive to phonetic variability than the `large` model. More research is needed to probe the different Whisper models - beyond the model cards (cf. (Mitchell et al., 2019)) that are proposed on the Whisper github [7] - but the analysis of the `tiny` models transcriptions of L2 speech clearly has a future for ICALL systems.

Acknowledgements

This publication has emanated from research supported in part by a 2021 research equipment grant (PAPTAN project)[8] from the Scientific Platforms and Equipment Committee and by a King's College/Université Paris Cité joint funding (Deep Learning for Language Assessment), both under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

[7] https://github.com/openai/whisper/blob/main/model-card.md

[8] Plateforme pour l'apprentissage profond pour la traduction automatique neuronale, in English: Deep Learning for Machine Translation at Université Paris Cité . See the description of the platform on the project website: https://u-paris.fr/plateforme-paptan

References

Arthur S Abramson and Douglas H Whalen. 2017. Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63:75–86.

Eric Atwell, Peter 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.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alfonso Caramazza, Grace H Yeni-Komshian, Edgar B Zurif, and Ettore Carbone. 1973. The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, 54(2):421–428.

Vincent Chanethom and Alice Henderson. 2022. Alignment in ASR and L1 listeners' recognition of L2 learner speech: A replication study. In *15th International Conference on Native and Non-native Accents of English*, Łódź, Poland. Université de Łódź.

George Christodoulides. 2016. *Effects of cognitive load on speech production and perception*. Ph.D. thesis, UCL-Université Catholique de Louvain.

Jonathan Dalby and Diane Kewley-Port. 1999. Explicit pronunciation training using automatic speech recognition technology. *CALICO*, 16(3):425–445.

Christelle Exare. 2017. *Les aspirations intrusives dans l'anglais des apprenants francophones*. Ph.D. thesis, Université Sorbonne Nouvelle.

Christelle Exare. 2022. Awareness of glottal settings for the production of /h/-initial and vowel-initial words in french learners of l2 english. *Anglophonia*, 32.

Georgi Gerganov. 2003. whisper.cpp : A high-performance inference of OpenAI's whisper automatic speech recognition (asr) model.

James Howard II. 2020. Phonetic spelling algorithm implementations for R. *Journal of Statistical Software*, 25(8):1–21.

Mathilde Hutin, Adèle Jatteau, Ioana Vasilescu, Lori Lamel, and Martine Adda-Decker. 2020. Lénition et fortition des occlusives en coda finale dans deux langues romanes : le français et le roumain (lenition and fortition of word-final stops in two Romance languages: French and Romanian). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, pages 289–298, Nancy, France. ATALA et AFCP.

Chen W. Inceoglu, S. and H. Lim. 2023. Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, 35(1):89–104.

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.

Takeki Kamiyama and Maelle Amand. 2023. Perception of word stress amongst French learners of English: nuclear tone suffix. In *Proceedings of the 20th International Congress of Phonetic Sciences, Prague 2023*, ID: 43, pages 2383–2387.

Roger Kingdon. 1958. *The Groundwork of English stress*. Longmans.

V. Krunic, G. Salvi, A. Bernardino, L. Montesano, and J. Santos-Victor. 2009. Affordance based word-to-meaning association. In *2009 IEEE International Conference on Robotics and Automation*, pages 4138–4143.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Leigh Lisker and Arthur S Abramson. 1967. Some effects of context on voice onset time in English stops. *Language and speech*, 10(1):1–28.

Courtney Mansfield. 2021. *ASR and human recognition errors: Predictability and lexical factors*. Ph.D. thesis, University of Washington.

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.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition

toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

C.L. Rogers, J. M. Dalby, and G. DeVane. 1994. Intelligibility training for foreign-accented speech: A preliminary study. *JASA*, 96(5):3348.

R Core Team. 2023. R: A language and environment for statistical computing.

Steven Weinberger. 2015. Speech accent archive. George Mason University. *http://accent.gmu.edu*.

Martijn Wieling, John Nerbonne, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa, and R Harald Baayen. 2014. A cognitively grounded measure of pronunciation distance. *PloS one*, 9(1):e75734.

Appendix

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. Their bloodless, pinched, yellow faces were like the faces of bilious invalids. 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.