Probing Whisper Predictions for French, English and Persian Transcriptions

Nicolas Ballier[1], Léa Burin[2], Behnoosh Namdarzadeh[2], Sara Ng[3], Richard Wright[3], Jean-Baptiste Yunès[4]
[1]LLF & CLILLAC-ARP, Université Paris Cité, F-75013 Paris, France
[2]CLILLAC-ARP, Université Paris Cité, F-75013 Paris, France
[3]University of Washington, USA
[4]IRIF, Université Paris Cité, F-75013 Paris, France
Contact: nicolas.ballier@u-paris.fr

Abstract

Whisper is a widely-used open-access Large Language Model (LLM) trained using a multilingual paradigm. As such it represents an important opportunity for researchers to study how multilingual LLMs function across languages. In this paper, we analyse Whisper's Large and Medium models for Persian, English and French using a transcription task. To investigate the calibration of Whisper models, we use a customised C++ version of Whisper to probe Whisper's internal representations by extracting the subtoken probabilities for transcriptions of speech samples of the target languages. We discuss our subtoken-based evaluation of prediction accuracy as a proxy for standard Word Error Rate evaluation of the different Whisper models. The accuracy of the ASR predictions is investigated as a function of target language and part of speech. Our analysis reveals an architectural bias for French and discrepancies in accuracy in relation to the size of the training data. The results of our novel subtoken-based evaluation supplement previously-reported cross-lingual evaluations of Whisper, and enable better fine-tuning by suggesting types of data that may improve calibration.

1 Introduction

Large Language Models (LLMs) are still perceived as black boxes. Recent papers have mostly described new state-of-the-art performance on transcription tasks with LLMs, but the reliability of different implementations has not, to the best of our knowledge, been investigated on the basis of the probability of the subtokens predicted by the LLMs. It should be borne in mind that LLMs do not predict tokens but subwords or subtokens, as the result of the byte pair encoding (BPE) (Sennrich et al., 2016), a compression algorithm adapted from Gage (1994). Taking advantage of its publicly available models and of the C++ implementation (Gerganov, 2003), whisper.cpp (hereafter "Whisper"), we probed the Whisper system and retroconverted the timestamps into a TextGrid (see Figure 1) in order to inspect the speech data. Our reverse engineering strategy is illustrated in Figure 1. We extracted timestamps and subtoken probability for each subtoken prediction.

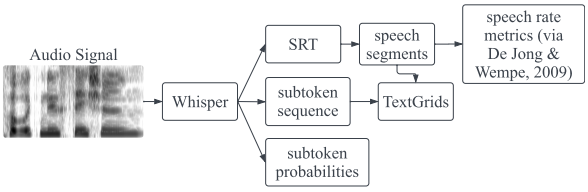

Figure 1: Extracting information from the Whisper pipeline (Radford et al., 2023) with Gerganov (2003).

Though Whisper has been trained with 680,000 hours of audio, out of which 117,000 hours represent 96 other languages than English, the distribution of the training data is heavily skewed, as indicated in the appendix of Radford et al. (2023)[1]. For inclusive spoken language science and technology, this discrepancy in the training data, as illustrated in Table 1, has a price worth investigation. We will show that Whisper's Large model has different calibration curves for Persian, French and English. The rest of the paper is structured as

Table 1: Number of hours of Whisper's training data for French, Persian and English, after Radford et al. (2023).

Speech Recognition	Translation
French : 9,752	French : 4,481
Persian : 24	Persian : 302
English : 563,000	

follows: Section 2 summarises previous research on Whisper and contextualises our contribution in this respect. Section 3 presents our experiment

[1]Since "Of those 680,000 hours of audio, 117,000 hours cover 96 other languages" we assumed that English was trained with 563,000 hours (680,000-117,000).

design, including the data and methodology. Section 4 presents the results. In Section 5 we discuss them. Section 6 concludes the paper.

2 Previous Research

Whisper is an audio Large Language model that has been trained for several tasks such as Voice Activity Detection, Transcription, textual translation into English and language detection (Radford et al., 2023). Less than two years after the public release of its models, more than 1,900 papers have been written using Whisper according to Google scholar. Many researchers have tried to optimise Whisper, for example by post-processing Whisper's outputs with LlaMa (Touvron et al., 2023) in a framework (Radhakrishnan et al., 2023) or to integrate Whisper in robots (Pande and Mishra, 2023). Whisper has been trained with LibriSpeech (Panayotov et al., 2015) data, whose features have been investigated for speech synthesis (Zen et al., 2019; Kakouros et al., 2023). Whisper has been tested with the FLEURS dataset (Conneau et al., 2023) for the Spoken Language Identification (SLID) task (Augenstein and Salaj, 2023) and used for Spoken Language Understanding (SLU) (Wang et al., 2023). Whisper has also been used for deep fake analysis (Kawa et al., 2023) and the whisper.cpp (Gerganov, 2003) implementation has been used to score second language speech (Ballier et al., 2023a). Coupled with an SVM classifier, Whisper showed good results when classifying vocal intensity categories (soft, normal, loud, and very loud) from speech signals (Kodali et al., 2023). Sun et al. (2023) have tested biasing lists to improve Whisper's speech recognition, which has also been improved when piped to the LLM LlaMa (Touvron et al., 2023) to select Whisper ASR outputs (Radhakrishnan et al., 2023). Analysing different varieties of English, Graham and Roll (2024) showed that Whisper performance was better for read speech than for spontaneous speech. They also showed that performance for Canadian and American English was comparable, but it was poorer for British and Australian English. Previous research on Whisper outputs has shown that the different segments produced by the different models are not identical in numbers and scope and differ from the speech signal (Ballier et al., 2023b). Several probing methods have been applied to LLMs, for example probing prompts (Qi et al., 2023), but Whisper probability distributions have not been investigated, to the best of our knowledge. The closest work to our research is a previous attempt to understand the information flux for the plural agreement in French, using a forced aligner and attention heatmaps, showing that agreement is dealt with in Whisper by the decoder, not the encoder of the Whisper `Transformer` architecture (Mohebbi et al., 2023).

3 Material and Methods

One previous study (Ballier et al., 2023b) suggests that when the Whisper outputs are not normalised (contrary to the normalisation procedure used in Radford et al. (2023)'s benchmarks and described in its appendix), word error rate (WER) is lower for the medium model than for the large model. We wanted to investigate the accuracy of the two models, as well as investigate how degraded the performance is when the training data size decreases. To this aim, we resorted to the calibration curve method, that plots the probability assigned to the subtoken (x) on the accuracy of its prediction (y axis). Best calibrated models are close to the $x = y$ axis and overconfident models are much above this axis. We present the calibration curve method before specifying the data we used for our tests.

3.1 The Calibration Curve Method

Assuming the probability assigned by the system to predicted subtokens is a correlate of confidence, we believe that for trustworthy AI we should investigate subtoken probabilities, especially when the prediction is wrong. A method to achieve this reliability analysis is the "calibration curve" method. This method has been used to analyse neural networks (Guo et al., 2017; Minderer et al., 2021) and, recently, to assess LLMs from a semantic point of view. For instance, Levinstein and Herrmann (2024) use calibration curves to assess the truthfulness of LLM statements on specific datasets and claims that "calibration provides another metric for evaluating the quality of the probes' forecasts"[2]. Calibration allows researchers to examine whether the model predictions are on average too certain (overconfident) or too uncertain (underconfident) (Minderer et al., 2021), this paving the way for LLM recalibration (Chen et al., 2024). Because our analysis is based on subtokens, we also computed a regression model to assess the role of subtokenisation, fitting a logistic model with the number of

[2]The concept was used initially to analyse the reliability of weather forecasts (Brier, 1950; DeGroot and Fienberg, 1983).

subtokens as one of the predictors.

3.2 Logistic Regression Modelling

We fitted a logistic regression model with accuracy (success rate) of the Whisper predicted subtokens as the predicted variable and several variables for predictors. We tested duration, segment, speaker, overlaps, speech rate and phonation rate. We extracted the turns from the official transcripts of the corpus (Branca-Rosoff, Sonia, 2013). We also computed the number of subtokens required to represent a token in the final transcription and POS-annotated the corresponding token. We used one of the treebanks for English to annotate the data from the ATAROS corpus (Freeman et al., 2014). We used the EWT model for the universal dependency annotation, based on English Web Treebank corpus (Silveira et al., 2014).

3.3 Data

We used challenging data to test the ASR task, since the LLM was trained with read speech from Librispeech (Panayotov et al., 2015).

- **Persian** We used two recordings from two female Tehrani speakers reading 120 sentences containing a dislocation. The sentences, extracted from various sources, reflect a formal register. Each sentence encapsulates the linguistic phenomenon of dislocation. Speakers recorded their voices on Zoom while reading each sentence aloud, since Whisper was trained to deal with noisy environments (Radford et al., 2023). We avoided overlaps and spontaneous speech because of the number of hallucinations observed when transcribing Persian.

- **French** For French, we used almost one hour (55 min.) of spontaneous French collected for the CFPP reference corpus (corpus de français parlé parisien) (Branca-Rosoff et al., 2009). This conversation of a dyad was collected in the early 2000's in Paris and has already been scrutinised from a prosodic perspective (Martin, 2020; Morel, 2011; Cresti et al., 2011).

- **English** For English, we used the ATAROS corpus (Freeman et al., 2014), designed to investigate stance and engagement in collaborative tasks. This corpus consists of dyadic conversations between unfamiliar interlocutors. Dyads of native English speakers from the Pacific Northwest of the United States (unknown to each other but roughly matched for age) completed a variety of collaborative tasks (Freeman, 2015). We present results from 2 sessions (56 min.) of mixed-gender dyads. We apply a temporal filter to the audio based on timings from human transcriptions of the target speaker, to mitigate non-target speech.

3.4 Data Extraction and Processing

We applied the following pipeline to our data:

- For reference corpora, we extracted the timestamps delimiting turns to create a speaker variable, and an overlap variable;

- With whisper.cpp, we extracted timestamps, subtoken predictions and the probability associated with each prediction;

- We qualitatively annotated the prediction of the Whisper model, assigning 0 to error and 1 to predictions. We report accuracy (success rate) and do not take into account omissions or word error rate (WER) because our analysis is at a subtoken level (we discuss the implications in relation to standard ASR based on WER in subsection 5.4);

- From whisper.cpp, we extracted the .SRT files that gave the timestamps of the segments created by the different Whisper models;

- With a series of scripts we computed the speech rate for each segment.

Using the C++ Whisper implementation (whisper.cpp), we also retrieved timestamps aligned to the Whisper segments, encapsulated in the extracted .SRT files. We then extracted the segment ID timestamps from the .SRT file, and mapped them onto the 16,131 prediction timestamps. We associated the 1,415 segments to their speech rates, computed with the De Jong and Wempe Praat plugin (De Jong and Wempe, 2009). We then checked for overlaps using the official transcript of the corpus (Branca-Rosoff, Sonia, 2013). Using the .trs (xml) file, we coded Whisper subtoken predictions corresponding to overlaps.

By default, we considered that we had no gold standard reference for the special tokens, so we discarded the special tokens (which we assumed to be correct predictions by default) as well as punctuation. Homophones were counted as errors as

they did not semantically match the reference transcription '*voir*' vs '*voire*'; '*m'aime*' vs '*même*'; '*ah c'est bon*' vs '*assez bons*'. Because Whisper was assessed with a normalisation procedure, we counted as correct '*17 and a half*' when the reference transcription had '*17,5*'.

3.5 A Brief Presentation of Whisper Byte-Pair-Encoding

For all languages and models, Whisper codes linguistic input as a composition of 51,866 subtokens. We provide a series of examples of the different types of tokens acknowledged in the HuggingFace documentation[3] and which can be explored in the dictionary of subtokens. We indicate the subtoken ID (number) which we found in the dictionary of subtokens.

- 50,255 linguistic subtokens, corresponding to English words or fragments for French or graphemes for languages like Persian;
- special tokens, some of them corresponding to boundaries of the `Transformer`: the end of text and end of sentence subtokens 50257 *[_EOT_]* and 50258 *[_SOT_]*;
- 100 extra-tokens labelled *[_extra_token_50259]* to *[_extra_token_50359]*;
- 7 special tokens are also acknowledged in the literature such as 50360 *[_SOLM_]*, 50361 *[_PREV_]*, 50362 *[_NOSP_]*, 50363 *[_NOT_]* and 50364 *[_BEG_]*. *[_BEG_]* corresponds to the beginning of the 30 second window when the sound file is processed by Whisper;
- 1,500 out-of-vocabulary OOV subtokens from *[_TT_1]* to *[_TT_1500]*. We will show that they correspond to temporal subtokens and we examine their status in subsection 4.2.

Our pipelines to investigate the Whisper inner computations is available on the GitHub of the sixth author.[4] We created automated scripts with R for the transformation of Whisper outputs into Praat TextGrids.

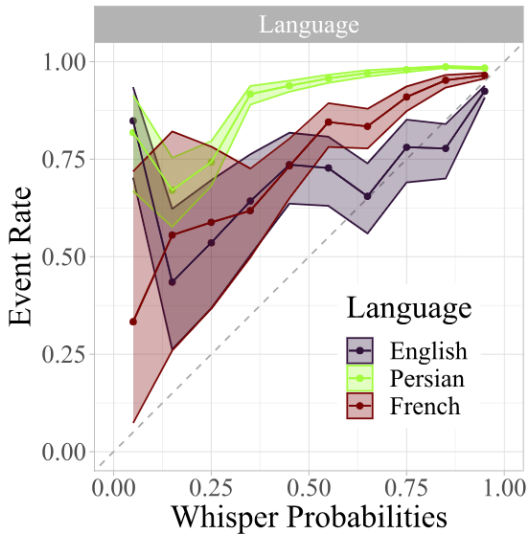

Figure 2: Calibration Curve for the Transcription of English, French, and Persian.

4 Results

4.1 Language Effects

We compared the calibration curves for the large models for the transcription of the three languages. Figure 2 shows the overconfidence of the Whisper model for Persian and French, well above the x=y line corresponding to the ideal calibration. When transcribing English, the predictions of the large model only partially overlap with ideal calibration.

4.2 Whisper's Internal Correlates to Temporal Values

One of the in-built limitations of the Whisper architecture is that audio inputs are limited to 30 second segments. When audio duration is greater than 30 seconds, the model must additionally truncate the audio at intermediate intervals. It appears that the so-called *TT* tokens may be outputs from this process. We analysed the main outputs of out of vocabulary *TT* tokens predicted by Whisper every time a punctuation symbol was used. We also analysed the property of the out of vocabulary tokens, the special tokens corresponding to end of text, end of sentence, and BEG, which structures the windowing of Whisper. In this subsection, we present the different types of results we obtained based on Whisper medium outputs on the Inventory

[3] https://huggingface.co/docs/transformers/model_doc/whisper

[4] https://github.com/jbyunes/whisper.cpp

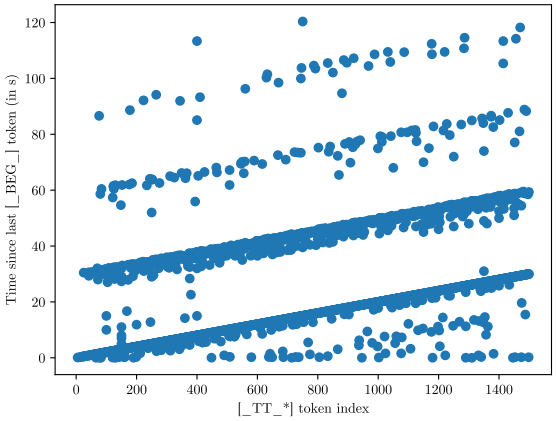

Figure 3: Comparision between the indices of *[_TT_*]* tokens and the time since the previous *[_BEG_]* token, across audio data from the ATAROS corpus.

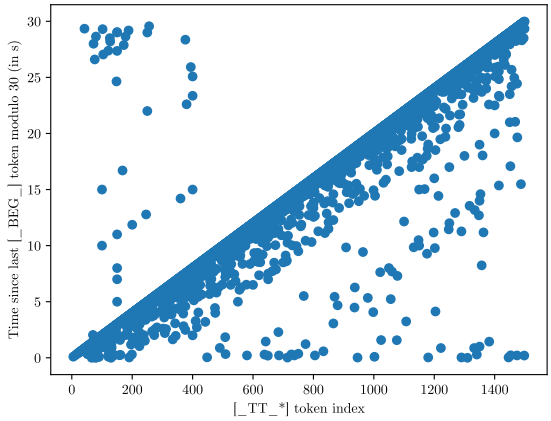

Figure 4: Comparison between the indices of *[_TT_*]* tokens and modulated spurt time, across audio data from the ATAROS corpus.

and Budget subtasks of the ATAROS dataset (Freeman et al., 2014). Based on the hypothesis that the numeric index of the *[_TT_*]* tokens were approximate to 20ms intervals since start of an audio span, we conducted a linear regression between the index of all *[_TT_*]* tokens in the output and the reported time difference between the token and the nearest *[BEG]* token in the previous output. Based on the observation that in some regions of the output the *[_TT_*]* token indices seem to "reset" without an intervening [BEG] token, we also conduct a linear regression between the *[_TT_*]* token indices and the time since the previous *[_BEG_]* token ***modulo 30s***. Figure 3 compares the token indices to the time since the previous *[_BEG_]* token, and Figure 4 compares the token indices to the modulated time since the previous *[_BEG_]* token. The regressions for both settings were significant ($p < 1e-15$). The r-value for the regression between the token indices and the raw time since the *[_BEG_]* token was 0.771, and when comparing to modulated time the r-value was 0.990.

4.3 Architectural Bias

Because Whisper predicts subtokens, not tokens, after the byte pair encoding (Sennrich et al., 2016), we created a subtoken_cnt variable corresponding to the number of subtokens needed to represent a given token. Previous research on neural machine translation has shown that gender bias for French into English translations can be sensitive to the number of subtokens required to represent tokens referring to female occupational nouns. We observed a similar architectural bias for French since the accuracy decreases with the number of subtokens, as can be seen on Figure 5. More research is needed to analyse how this might be a confounding factor for the mistranscription of named entities. The architectural bias was not observed in the regression model for Persian but it may be because the dataset was too small. Architectural bias (if not training bias) was observed in the accuracy of the prediction of the subtokens, which was distributed unevenly. Some subtokens were systematically accurate or wrong in the transcriptions, independently of the estimated probability as can be seen on Fig-

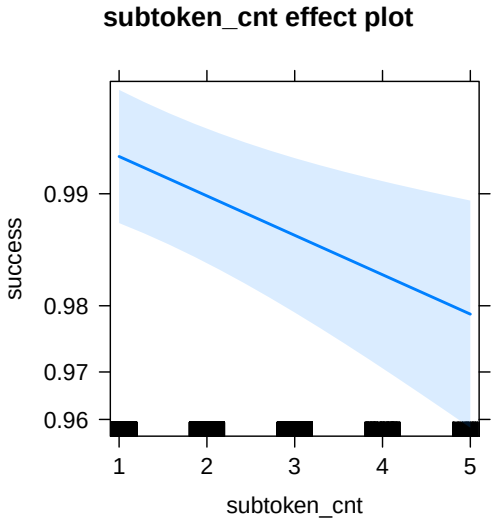

Figure 5: Effect of the number of subtokens on the transcription task for French (Large model).

ure 6.[5]

4.4 POS Effect

Universal dependency annotation provides two types of part-of-speech annotation, one which is more general and follows universal guidelines to describe categories that are posited to be universally valid for all languages (upos). Other language-related tagsets entrenched in the metalinguistic tradition of a given language (xpos) are also acknowledged, as is the case here for English with the Penn Treebank tagset displayed in Table 2. The default tagset (WET) used to analyze English was trained on web data, so it was not particularly designed to annotate spoken corpora. Some observations can still be made, some part-of-speech categories being more prone to phonetic variation such as *to* (xpos: TO) and determiners (xpos:DT), which are more likely to undergo alternations between reduced and full vowels, as can be seen in Table 2. If we take into account the universal part-of-speech (upos) tagset, perfect success rates can be achieved for categories such as conjunctions of coordination. On the other hand, weak forms or determiners are likely to undergo more ambiguous transcriptions probably because of the weak forms of *a* and other determiners. Similarly, weak forms may account for the rather poor score for auxiliaries and pronouns. There is thus an effect of weak forms and their potential ambiguities.

Table 2: Best categories predicted.

xpos	n	success
TO	39	0.82
DT	102	0.90
PRP	68	0.94
WDT	15	1
CD	17	1
CC	66	1
IN	77	1
upos	**n**	**success**
AUX	71	0.95
PRON	127	0.96
NOUN	172	0.97
NUM	17	1
SCONJ	48	1
ADP	56	1
CCONJ	66	1

5 Discussion

5.1 Suggestions for Fine-tuning Whisper for Persian

It seems that the models for Persian are less robust, as a very strong inter-speaker variability can be observed. For the speech recognition of the same sentences, the success rate varies from 95% to 55% and this speaker effect can be seen on the calibration curves on Figure 7. Overfitting of the Persian model with Arabic data needs to be stressed as well. This can be explained by a partly commonly shared alphabet between the two languages. Persian has a few specific graphemes for the voiced velar and significantly more homophones than in Arabic. Nevertheless, some very specific Arabic letters are used instead of Persian like the nasal consonant, the alveolar nasal and other substitutions can be observed. Furthermore, the Perso-Arabic script used to write Persian is cursive, meaning that letters tend to have different shapes depending on whether they join with adjacent letters or not. The different graphotactics of Persian for initial, medial, and final characters are not represented in the sub-tokenization of Persian transcription. Considering the phonotactic and graphotactic constraints of Persian showcased in the transcription by Whisper, fine-tuning Whisper could be a way to improve the transcriptions of a language with low training data.

5.2 The Locus of Hallucinations

In the case of Persian, smaller models of Whisper exhibited some hallucinations, which can be attributed to the subtoken dictionary. However, these hallucinations were not present in the larger model. The occurrence of hallucinations is not consistent across different models. Specifically, in smaller models like the `small` model, numbers read by the speaker at the beginning of each sentence were often hallucinated. For English and French, we mostly observed “coda" hallucinations as in Figure 8. Within the two seconds after the end of speech intervals, transcriptions are provided in spite of the absence of speech signal. Our hypothesis is this comes from the training data (probably from

[5]This subtoken-by-subtoken analysis could be replicated with English and French but with more difficulty, as the order magnitude for the number of subtokens is 1 to 8 for Persian to French and 1 to 40 for Persian to English, an estimation based on Google’s Compact Language Detector 3.

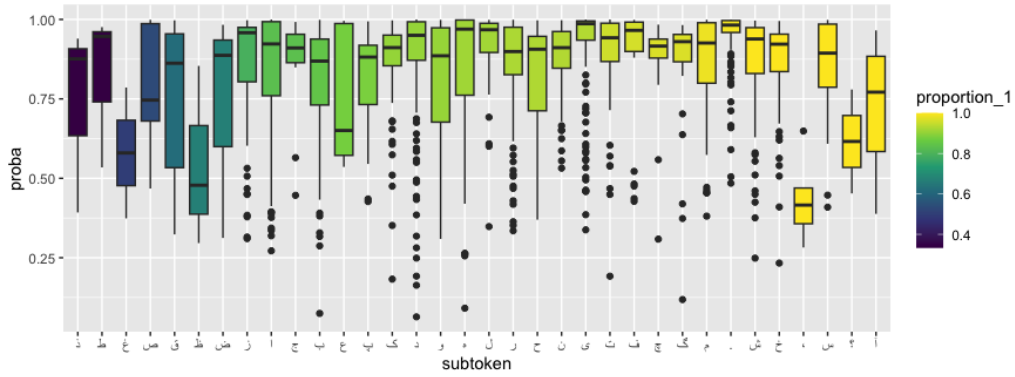

Figure 6: Distribution of probability for Persian subtokens of one character.

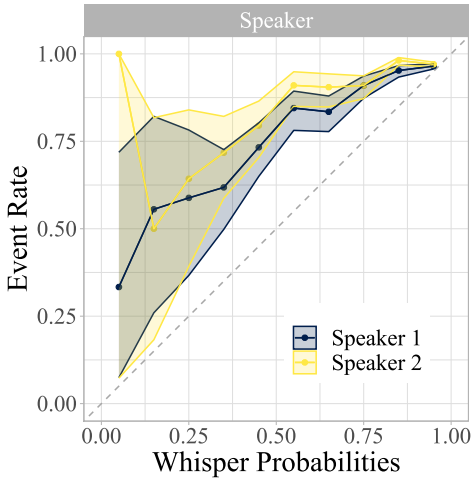

Figure 7: Speaker effect on the quality of the prediction in relation to the confidence of the model (Persian data, Large model).

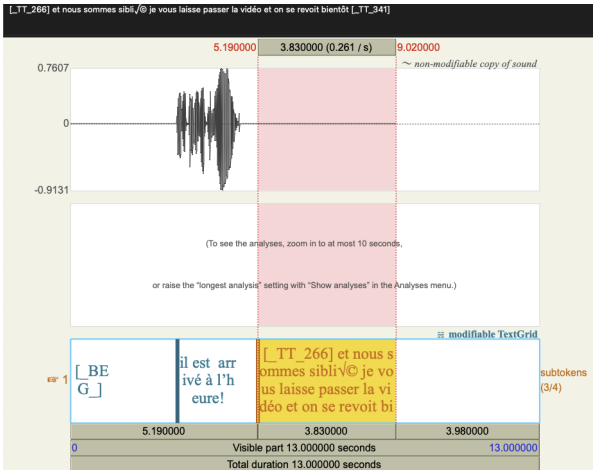

Figure 8: Coda hallucination in French. The hallucination disappears with the same Whisper model if the pause after the utterance is reduced.

YouTube) which contains final formulae like “see you soon", here “je vous laisse la vidéo et on se voit bientôt" (“I'll leave you the video and see you soon."). For English we often found “Thank you" in the coda hallucinations.

5.3 Reliability of Whisper's Timestamps?

Many special tokens separating subtokens have no duration and some .SRT files are uneasily retro-converted to TextGrids. Moreover, the timings do not match up very well with the word-level timings for ATAROS, which is why we reported two estimations for overlap labels – one based on Whisper's timing, and one version based on the human-aligned timings. Figure 9 shows the discrepancies of duration according to the two methods, whether for words or subtokens.

5.4 The Censorship of Repetitions

Our analysis of the success rate is a precision analysis rather than an analysis of recall. We based our analysis on the Whisper predictions, not on the official transcripts of the corpora when available. For English, we also computed an analysis of recall, namely comparing the Whisper predictions to the original transcription of the ATAROS data. As part of the discussion, we computed the difference between using reference text to the corpus as the baseline to which we annotated the prediction of the Whisper models, and we compared this method with the raw output of the Whisper models that was annotated only on the basis of the predictions. Using the first method, we report a 79% success rate, and then we re-aligned only the prediction of the LLMs and computed the success rate. In our accuracy-based analysis, the omissions from the scripts, and in particular all the censorship of the repetitions of the data, were more favourable to the

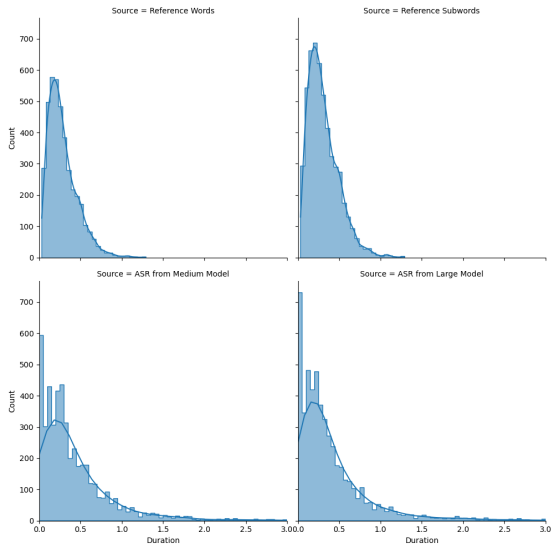

Figure 9: Distribution of duration according to the ATAROS reference transcription (top) and according to Whisper's Large model (ASR, bottom)

interpretation of the Whisper success rates, since we achieved 89% of success using this methodology based on the analysis of the Whisper output only. Among the Whisper omissions in the transcriptions, repetitions accounted for 11.9% of the omitted tokens. Numbers (and generally speaking, counting) accounted for 37.5% of the omissions from the reference text.

6 Conclusions

In this paper, we have probed Whisper outputs using a C++ implementation of Whisper (Gerganov, 2003) to probe the accuracy of transcriptions on a subtoken basis. We use subtoken probabilities and internally produced timestamps. We used reverse engineering to translate the inner working of a large language model, namely its prediction properties, to realign them to the speech signal with Praat (Boersma, 2024) TextGrids. Our method suggests coherent meanings to the special temporal subtokens *[_TT_*]* used by Whisper. This type of research aims to contribute to the explainability of LLMs. The same method could be applied to the translation task; we have here investigated the probabilities associated with the subtokens produced by Whisper for the transcription task. Estimating LLM ASR output at subtokens level allows us to access transcriptions at a finer-grained level and it paves the way for other analyses currently used in the semantic analysis of LLMs such as grouping loss (Perez-Lebel et al., 2023). It should also be noted that analysing subtokens is another way to ensure hallucination detection: subtokens representing Arabic or Japanese were observed for Persian. An unexpected finding is that Whisper scores only report the Persian letters in their isolated forms (abstract representation) and positional variants of letters as observed in the Whisper textual transcriptions seem to be the result of some post-processing. Further studies are needed to investigate this point.

We have shown the effect of size in the training effect, but also an architectural bias for French. It would be interesting to apply the same methodology to explore the probabilities assigned to the translation task to confirm these biases and effects. Analysing the Whisper performances on other languages may confirm one of our observations. With the R (R Core Team, 2024) package Calibratr (Schwarz and Heider, 2019), we also computed the Expected Calibration Error (ECE), which returns the maximum calibration error for equal-frequency binning model (Naeini et al., 2015) for the transcriptions (large model) of the three languages. With the proviso that we have only analysed the transcriptions of three languages with Whisper, a linear model can be fitted with the log of the size of the training data (adjusted R-square 0.99) and it may be the case that the ECE is inversely proportional to the log of the size of the training data, as can be observed on Figure 10.

Figure 10: According to Whisper's Large model (ASR task), the effect of the size of the training data

Limitations

As for Persian, our test set consists of read speech focusing on a linguistic construction, dislocation.

We have not used the fine-tuned XLSR-53 large model for speech recognition in Persian (Grosman, 2021) in this study. Using the train and validation splits of Common Voice 6.1 in this fine-tuned model may change the results.

References

Benedikt Augenstein and Darjan Salaj. 2023. Exploiting foundation models for spoken language identification. In *Proceedings of the LWDA 2023 Workshops: BIA, DB, IR, KDML and WM*, pages 28–40.

Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023a. 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.

Nicolas Ballier, Behnoosh Namdarzadeh, Maria Zimina, and Jean-Baptiste Yunès. 2023b. Translating dislocations or parentheticals : Investigating the role of prosodic boundaries for spoken language translation of French into English. In *Proceedings of Machine Translation Summit XIX: Users Track*, pages 119–132, Virtual. Association for Machine Translation in the Americas.

Paul Boersma. 2024. Praat: doing phonetics by computer. *http://www. praat. org/*.

Sonia Branca-Rosoff, Serge Fleury, Florence Lefeuvre, and Mat Pires. 2009. Discours sur la ville. corpus de français parlé parisien des années 2000.

Branca-Rosoff, Sonia. 2013. Entretien de anita musso (annotations).

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Lihu Chen, Alexandre Perez-Lebel, Fabian M Suchanek, and Gaël Varoquaux. 2024. Reconfidencing LLMs from the Grouping Loss Perspective. *arXiv preprint arXiv:2402.04957*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Emanuela Cresti, Massimo Moneglia, and Ida Tucci. 2011. Annotation de l'entretien d'Anita Musso selon la théorie de la langue en acte. *Langue française*, 170(2):95–110.

Nivja H De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Valerie Freeman. 2015. *The Phonetics of Stance-taking*. Ph.d. thesis, University of Washington, Seattle, USA.

Valerie Freeman, Julian Chan, Gina-Anne Levow, Richard Wright, Mari Ostendorf, and Victoria Zayats. 2014. Manipulating stance and involvement using collaborative tasks: an exploratory comparison. In *Proc. Interspeech 2014*, pages 303–307.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Georgi Gerganov. 2003. whisper.cpp : A high-performance inference of OpenAI's whisper automatic speech recognition (ASR) model. *https://github.com/ggerganov/whisper.cpp*.

Calbert Graham and Nathan Roll. 2024. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in Persian. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-persian`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Sofoklis Kakouros, Juraj Šimko, Martti Vainio, and Antti Suni. 2023. Investigating the Utility of Surprisal from Large Language Models for Speech Synthesis Prosody. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 127–133.

Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szymański, and Piotr Syga. 2023. Improved DeepFake Detection Using Whisper Features. In *Proc. INTERSPEECH 2023*, pages 4009–4013.

Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. 2023. Classification of Vocal Intensity Category from Speech using the Wav2vec2 and Whisper Embeddings. In *Proc. INTERSPEECH 2023*, pages 4134–4138.

Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Philippe Martin. 2020. L'annotation prosodique dans ORFÉO. *Langages*, 219(3):103–115.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260.

Mary-Annick Morel. 2011. Les paragraphes intonatifs d'Anita Musso: entre consensus coénonciatif et égocentrage colocutif. *Langue française*, 170(2):111–126.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, page 2901–2907.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Akshara Pande and Deepti Mishra. 2023. The synergy between a humanoid robot and whisper: Bridging a gap in education. *Electronics*, 12(19).

Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. 2023. Beyond calibration: estimating the grouping loss of modern neural networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. 2023. What is the limitation of multimodal llms? a deeper look into multimodal llms through prompt probing. *Information Processing & Management*, 60(6):103510.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023. Whispering LLaMA: A Cross-Modal Generative Error Correction Framework for Speech Recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016.

Johanna Schwarz and Dominik Heider. 2019. GUESS: Projecting machine learning scores to well-calibrated probability estimates for clinical decision making. *Bioinformatics*, 35(14):2458–2465.

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.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for English. In *LREC*, pages 2897–2904. Citeseer.

Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C. Woodland. 2023. Can Contextual Biasing Remain Effective with Whisper and GPT-2? In *Proc. INTERSPEECH 2023*, pages 1289–1293.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. 2023. WhiSLU: End-to-End Spoken Language Understanding with Whisper. In *Proc. INTERSPEECH 2023*, pages 770–774.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.