Automatic Speech Recognition for Non-native Speakers

Adélaïde Couplet[1], Anaïs Tack[2], and Anne-Catherine Simon[3]

[1] NADI Institute - University of Namur - Faculty of Computer Science - Belgium
[2] ITEC - Katholieke Universiteit Leuven Campus Kortrijk - Belgium
[3] Valibel - UCLouvain - Louvain-la-Neuve - Belgium

Keywords: Natural Language Processing · Automatic Speech Recognition · Non-native Speaker · Whisper

1 Introduction and Research Context

Automatic speech recognition (ASR) is a subfield of natural language processing that aims to automatically transcribe spoken language into text. In this work, we focus on the challenges of transcribing the speech of non-native speakers. With the advent of transformer-based architectures [8], ASR performance has significantly improved under clean conditions. However, when the data is noisy or "non-standard", for example, when the speaker has an accent or a speech disorder, ASR systems still exhibit limited robustness [1]. Improving ASR for such non-standard speech has the potential to support second language learning.

Previous studies addressed ASR for non-native speech in various ways. For pedagogical purposes, Li et al. used ASR to detect errors and help students improve their pronunciation [5]. Inceoglu et al. compared human comprehension with ASR performance on non-native speech [4]. Researchers also investigated model adaptation specifically for non-native productions [1, 3]. More closely related to our work, Ballier et al. evaluated Whisper's ability to identify recurrent pronunciation errors in English learners' speech [2]. Our study extends this line of research by examining French productions of Dutch-speaking secondary school students, an audience that has received little attention in research [6].

2 Methodology

We conduct this research using a corpus provided by Dr. Ann-Sophie Noreillie [6] at KU Leuven in Kortrijk. The corpus consists of recordings of Dutch-speaking secondary school students with a B1 CEFR level in French, as well as recordings of native French-speaking students. Both groups were asked the same set of questions on two topics: a medical appointment and a job interview. In our thesis, the productions of the native French speakers are used for comparison.

Transcriptions are obtained using Whisper, OpenAI's model for ASR [7]. There are different versions of Whisper, varying in the amount of training data and the number of layers. We test six versions in this study: *tiny*, *base*, *small*, *medium*, *large*, and *large-v2*. We transcribe all data with all six versions of the

model. We then compute the Word Error Rate (WER) and generate alignment representations, which also enable a qualitative error analysis.

This work addresses the following three research questions: RQ1 - Which version of Whisper provides the best results for transcribing the dialogues in Noreillie's corpus? RQ2 - Do Whisper models perform less effectively on non-native speech compared to native speech? RQ3 - Which version of Whisper achieves the best results specifically for non-native speakers' speech?

3 Experiments and Results

We first analyze the results for all speakers combined (native and non-native). In general, larger models (i.e., models with more layers and trained on more data) achieve better performance. However, this initial analysis reveals that the obtained WER values are very high compared to state-of-the-art results on other corpora [7]. At first, we assumed this was solely due to the presence of non-native speakers. However, further analysis of the natives' productions shows that additional factors contribute to the high error rates. Specifically, mismatches between our gold transcriptions and Whisper's output introduce systematic errors. For example, Whisper automatically adds capitalization and punctuation, whereas our reference transcriptions do not. Moreover, because the recordings are dialogues, we encounter issues related to speaker diarization (attributing a speaking slot to one of the speakers). Since Whisper does not natively handle diarization, the presence of two speakers in each recording creates additional transcription errors.

WER	Tiny	Base	Small	Medium	Large	Large-v2
NS	0.76	0.68	0.59	0.58	0.57	0.58
NNS	0.94	0.72	0.62	0.59	0.59	0.57

Table 1. WER for native speakers (NS) and non-native speakers (NNS)

Despite the relatively high WER values, the *medium*, *large*, and *large-v2* models consistently outperform the others, thus answering RQ1. When comparing results across speaker groups, Table 1 shows that smaller models (*tiny*, *base*, *small*) perform worse on non-native speech than on native speech. By contrast, the larger models (*medium*, *large*, *large-v2*) yield almost identical WER scores for both groups, thereby addressing RQ2. With respect to RQ3, WER results alone suggest that the *medium*, *large*, and *large-v2* models are the most suitable. Among them, *medium* represents an attractive compromise, being faster and less resource-intensive. However, our qualitative analysis on a sample of the corpus reveals an important limitation: the larger Whisper models tend to "overcorrect" learners' productions. Instead of faithfully transcribing what was said, the models sometimes normalize or reformulate utterances to produce grammatically correct sentences. In certain cases, this alters the meaning by adding or modifying words. For applications in language learning, where capturing learners' errors is crucial, this behavior suggests that Whisper, despite its strong performance in WER, may not be the most appropriate tool for pedagogical transcription tasks.

References

1. Bada, I., Fohr, D., & Illina, I. (2020). Reconnaissance automatique de la parole: génération des prononciations non natives pour l'enrichissement du lexique. In 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 (pp. 27-35). ATALA; AFCP.
2. Ballier, N., Méli, A., Amand, M., & Yunès, J. B. (2023, December). Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English. In 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023) (Vol. 6, No. 282-292).
3. Bouselmi, G. (2008). Contributions à la reconnaissance automatique de la parole non-native. Theses, Université Henri Poincaré - Nancy I.
4. Inceoglu, S., Chen, W. H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. ReCALL, 35(1), 89-104.
5. Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. (2016, March). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6135-6139). IEEE.
6. Noreillie, A.-S. (2019). It's all about words. Three empirical studies into the role of lexical knowledge and use in French listening and speaking tasks. (Unpublished doctoral dissertation). PhD thesis, KU Leuven Campus Antwerpen.
7. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.