The Impact of Foreign Accents on the Performance of Whisper Family Models Using Medical Speech in Polish

Szymon Zaporowski
Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland *szyzapor@pg.edu.pl*

Abstract

The article presents preliminary experiments investigating the impact of accent on the performance of the Whisper automatic speech recognition (ASR) system, specifically for the Polish language and medical data. The literature review revealed a scarcity of studies on the influence of accents on speech recognition systems in Polish, especially concerning medical terminology. The experiments involved voice cloning of selected individuals and adding prosodic contours with Russian and German accents, followed by transcription of these samples using all available models from the Whisper family and comparison with the original transcription. The results of these initial experiments suggest that the Whisper model struggles with foreign accents in the context of Polish language and medical terminology. This highlights the need for further research aimed at improving ASR systems for foreign accents and medical terminology.

Keywords: Automatic Speech Recognition, Whisper, Medical Language Recognition, Speech Processing

1. Introduction

Over the past few years, significant advancements have been made in the field of automatic speech recognition. Large models based on artificial neural networks and transformer architectures, such as Wav2Vec, Whisper, and Canary, have enabled high-quality transcription for numerous languages [2, 3], [7] . However, a persistent challenge remains in how these models perform with speech that deviates from the standard input of training datasets, particularly speech with foreign accents. This issue remains a central topic of ongoing research, even with the introduction of advanced models such as Whisper [4]. Most of these studies pertain to the English language, which historically encompasses a broad array of accents used by tens of millions of people worldwide. In the context of the Polish language, regional dialects are most common, along with accents derived from the so-called "Eastern Wall" of Poland. Due to geopolitical shifts, there is an increasing presence on the streets of Poland of individuals speaking Polish with Eastern accents, such as Ukrainian or Russian. This also applies to doctors from Ukraine, of whom, according to data from the Polish Ministry of Health, there are nearly 4,000 [6].

Particularly in medical contexts, where precise communication is crucial, these transcription systems could potentially malfunction when encountering foreign accents. This raises the necessity for specific research to verify the extent of this issue and to determine its impact on the accuracy of medical documentation. It is essential to assess whether current speech recognition technologies are adequately equipped to handle such variations in speech and to explore potential solutions or adjustments to improve recognition accuracy for accents not typically included in standard training sets.

The objective of the experiment outlined in this article is to evaluate the performance of a range of models from the Whisper family specifically for medical Polish language, using data that includes foreign accents.

2. Related works

In the literature, there are several studies focused on the topic of medical speech recognition in Polish. These studies primarily concentrate on the accurate recognition of speech using ASR (Automatic Speech Recognition) models provided through APIs by companies such as Google, Microsoft, and employing the Whisper model [5], [10]. A portion of the research also addresses the impact of accents on speech recognition; however, these studies pertain to the English language, for instance, the impact of a Polish accent on English speech recognition [8, 9].

During the literature review, no studies were found that specifically address the influence of accents on speech recognition for the Polish language, especially considering medical terminology. This indicates a significant research gap, suggesting the need for investigations into how different accents in Polish language affect the performance of speech recognition systems when processing specialized medical vocabulary.

3. Methodology

For the experiment, all models from the OpenAI Whisper family were utilized, ranging from tiny to large-v3. Fragments from the CommonVoice dataset served as the benchmark dataset for the Polish language [1]. Ten speakers were selected, with ten samples for each individual. Each sample lasted between 5 to 15 seconds. Subsequently, each sample underwent a voice cloning process using the IMS Toucan framework, with a foreign accent added. The selected accents were Russian and German. The next step involved transcribing each of the resulting samples using the Whisper family models and comparing the results with the original transcription. Additionally, normalization techniques developed by the creators of Whisper were integrated into the pipeline to achieve results as close as possible to those obtained by OpenAI. For evaluating the quality of the transcriptions, standard metrics supported by the Jiwer library were employed. These included: Word Error Rate (WER), Match Error Rate (MER), Character Error Rate (CER), Word Information Lost (WIL), Word Information Preserved (WIP). The obtained results for each person were averaged to yield a single metric for the entire tested dataset. In the results section, the metrics WER, MER, and CER are presented as the most significant.

This process generated benchmark results. The procedure was then repeated for selected recordings involving medical professionals. This data included 10 individuals. For each individual there were 10 recordings, varying in duration from 5 to 25 seconds (average sample length was 14.8 second). These recordings were captured in controlled studio environments using professional-grade equipment. The participants read texts that had been previously curated by experts from the medical community. The data employed were sourced from the Polish Medical Speech Corpus (PMS Corpus), which is currently under development as part of a research initiative aimed at training ASR models to accurately recognize speech from medical staff. The voices were again cloned using IMS Toucan with the selected accents, followed by the transcription process and comparison with the original text.

It should be noted that the conducted experiment is only a proof-of-concept and is intended to be carried out on thousands of recordings for hundreds of individuals.

4. Results and discussion

This section presents a comprehensive analysis of the performance of various Whisper models, ranging from Tiny to Large-v3, across three sets of data: a part of the CommonVoice dataset (CV) and Polish Medical Speech (PMS) data spoken in Polish with different accents - Polish, Russian, and German. The evaluation metrics used include WER, MER, CER. In Table 1, we present benchmark results for the original data from the CommonVoice corpus and the Polish Medical Speech (PMS) with a Polish accent.

For speech with a native Polish accent, the results indicate a steady improvement in all metrics from the Tiny to the Large-v3 model. The performance on the medical speech (PMS) dataset generally lags slightly behind the CommonVoice dataset, likely due to the specialized vocabulary and acoustic features of medical speech.

Table 1. Results for part of CommonVoice dataset (CV) and Polish Medical Speech (PMS for Polish speech with Polish accent

Model type/metrics	WER-PMS	WER-CV	MER -PMS	MER-CV	CER-PMS	CER-CV
Tiny	0.69	0.71	0.61	0.61	0.23	0.18
Base	0.60	0.65	0.54	0.57	0.19	0.15
Small	0.52	0.59	0.47	0.52	0.16	0.13
Medium	0.46	0.53	0.42	0.47	0.14	0.11
Large	0.41	0.49	0.38	0.43	0.13	0.10
Large-v2	0.38	0.46	0.35	0.41	0.12	0.09
Large-v3	0.35	0.44	0.33	0.39	0.11	0.09

Table 2 shows the results for Russian accent applied to data form CV and PMS corporas.

Table 2. Results for part of CommonVoice dataset (CV) and Polish Medical Speech (PMS) for Polish speech with Russian accent

Model type/metrics	WER-PMS	WER-CV	MER-CV	MER-PMS	CER-PMS	CER-CV
Tiny	0.81	0.78	0.67	0.71	0.30	0.22
Base	0.73	0.74	0.63	0.64	0.25	0.20
Small	0.64	0.69	0.60	0.58	0.22	0.18
Medium	0.58	0.65	0.56	0.52	0.19	0.16
Large	0.53	0.61	0.53	0.48	0.17	0.15
Large-v2	0.49	0.58	0.51	0.45	0.16	0.14
Large-v3	0.47	0.56	0.49	0.43	0.15	0.13

When evaluating the models on Polish speech with a Russian accent, there is a noticeable increase in error rates compared to the native accent, reflecting the additional challenge posed by accent variations. Despite this, the upward trend in performance with larger models persists, highlighting their robustness.

Table 3 presents results for German accent on both CV and PMS datasets.

Table 3. Results for part of CommonVoice dataset (CV) and Polish Medical Speech (PMS) for Polish speech with German accent

Model type/metrics	WER-PMS	WER-CV	MER -PMS	MER-CV	CER-PMS	CER-CV
Tiny	1.08	0.85	0.87	0.74	0.53	0.25
Base	0.93	0.80	0.79	0.71	0.41	0.24
Small	0.81	0.75	0.71	0.66	0.34	0.21
Medium	0.76	0.70	0.67	0.62	0.32	0.19
Large	0.68	0.66	0.60	0.58	0.28	0.17
Large-v2	0.63	0.63	0.57	0.56	0.26	0.16
Large-v3	0.60	0.61	0.54	0.54	0.24	0.15

The German accent presents the highest error rates across all models and metrics. However, like the other accents, the error rates decrease as the models increase in size. The large disparity in performance between the German-accented Polish Medical Speech

and the CommonVoice dataset emphasizes the difficulty of recognizing heavily accented speech in a specialized domain like medicine. The results elucidate several critical points. First, the presence of foreign accents significantly affects the performance of ASR systems, with the German accent having the most substantial impact. Second, the Whisper models demonstrate considerable capability in adapting to accent variations, although the performance is consistently better on generic datasets compared to specialized medical speech.

5. Conclusion

The experimental results underscore the challenges and opportunities in enhancing ASR systems for medical environments. Specifically, the impact of foreign accents on ASR accuracy highlights a critical area for further research. The findings suggest that future development efforts should focus on training models with a broader range of accent data and consider the customization of models specifically for medica\l contexts. By increasing the models' exposure to diverse accents and specialized medical vocabularies, it is feasible to substantially reduce the performance disparities observed between general and medical speech datasets. Such advancements could improve the reliability and utility of ASR systems in healthcare settings, ultimately contributing to better communication and documentation accuracy within medical practices.

Acknowledgments:

This research was supported by the Polish National Centre for Research and Development (NCBR) within the project: “ADMEDVOICE - Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process”. No. INFOSTRATEG4/0003/2022.

References

1. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: ..12th Int. Conf. on Language Resources and Evaluation, .pp. 4218–4222. (2020)
2. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst., pp.1–19 (2020)
3. Dhawan, K., Rekesh, Kd., Ginsburg, B.: Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer. In: Winata, G., Kar, S., Zhukova, M., Solorio, T., Diab, M., Sitaram, S., Choudhury, M., and Bali, K. (eds.) 6th Workshop on Computational Approaches to Linguistic Code-Switching pp. 74–82 (2023)
4. Graham, C., Roll, N.: Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits. JASA Express Lett. 4 (2), (2024)
5. Kuligowska, K., Stanusch, M., Koniew, M.: Challenges of Automatic Speech Recognition for medical interviews - research for Polish language. Procedia Comput. Sci. 225, pp. 1134–1141. (2023)
6. Puls Medycyny (2024), Prawie 4 tys. lekarzy z Ukrainy otrzymało zgodę na wykonywanie zawodu w Polsce, https://pulsmedycyny.pl/prawie-4-tys-lekarzy-z-ukrainy-otrzymalo-zgode-na-wykonywanie-zawodu-w-polsce-1209285, Accessed: ,
7. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning. JMLR.org (2023)
8. Radzikowski, K., Wang, L., Yoshie, O., Nowak, R.: Accent modification for speech recognition of non-native speakers using neural style transfer. EURASIP J. Audio, Speech, Music Process. 2021 (1), 11 (2021)
9. Trzeciakowska, J.: Non-Native English Speakers’ Attitudes Towards Polish-Accented English. Theor. Hist. Sci. 17, pp .65–69. (2020)
10. Zielonka, M., Krasiński, W., Nowak, J., Rośleń, P., Stopiński, J., Żak, M., Górski, F., Czyżewski, A.: A survey of automatic speech recognition deep models performance for Polish medical terms. In: 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 19–24. (2023)