Comparing Speech Synthesis Models for Polish Medical Speech Naturalness

Wiktor Krasiński
Gdańsk University of Technology
Gdańsk, Poland *s179987@student.pg.edu.pl*

Przemyslaw Rośleń
Gdańsk University of Technology
Gdańsk, Poland *s180150@student.pg.edu.pl*

Andrzej Czyżewski
Gdańsk University of Technology
Gdańsk, Poland *andcz@sound.eti.pg.gda.pl*

Marta Zielonka
Gdańsk University of Technology
Gdańsk, Poland *marta.zielonka@pg.edu.pl*

Abstract

This research investigates the perceived naturalness of synthesized speech in the context of Polish medical terminology, a critical factor for applications such as voice-enabled medical dialogue systems. We conducted a comparative analysis of three speech synthesis models: SpeechGen, ElevenLabs, and a version of ToucanTTS fine-tuned on a specialized corpus of Polish medical recordings. The evaluation employed objective measures, the NISQA metric, and subjective assessments through Mean Opinion Score (MOS) surveys. Our findings indicate that SpeechGen and ElevenLabs produce synthesized speech that closely rivals the naturalness of human speech, as evidenced by both NISQA scores and MOS ratings. In contrast, despite improvements, the fine-tuned ToucanTTS model did not achieve comparable levels of perceived naturalness. Notably, participants occasionally rated the advanced synthesized speech as more natural than human speech recorded in non-studio environments, underscoring the potential of these technologies in real-world applications. This study emphasizes the significance of naturalness in enhancing user experience, particularly in specialized linguistic domains. It provides insights into speech synthesis's current capabilities and limitations for less-resourced languages like Polish.

Keywords: Speech Synthesis Models, Polish Medical Speech, Speech Naturalness

1. Introduction

Recent advancements in deep learning have significantly improved speech synthesis, broadening its applications from virtual assistants to accessibility tools and creating a need for rigorous quality evaluation. This study focuses on assessing the quality of deep learning-based synthesizers, with an emphasis on naturalness as a critical factor in user experience. Synthesizing speech in Polish, particularly with medical vocabulary, presents unique challenges due to the language's complex phonetics and specialized jargon. As of 2023, Poland has approximately 163,200 licensed medical doctors residing in the country. According to various sources, the total number of Polish-speaking medical doctors globally can be estimated at 165,000 to 170,000. Therefore, Polish medical speech is an example of a minority language worth considering, especially given the ongoing project to develop voice-enabled medical documentation systems for

Polish-speaking doctors. It is crucial to understand how well modern synthesizers perform in this niche domain. Traditional synthesis techniques, such as concatenative and parametric methods, have given way to neural network-based models, including WaveNet [1], Tacotron [2], and FastSpeech [3], which produce more natural-sounding speech. This research aims to refine the ToucanTTS model [4] and evaluate its performance against leading systems, such as SpeechGen [5] and ElevenLabs [6], using both objective (NISQA) and subjective (MOS) quality metrics. By employing the Admedvoice dataset [7] , a repository of Polish medical recordings, this study seeks to advance speech synthesis in specialized fields and inform future developments in voice-enabled medical tools. We acknowledge that our study evaluates existing TTS systems rather than proposing novel models or techniques. However, we believe it provides a meaningful contribution by assessing TTS performance in the underrepresented domain of Polish medical speech—a niche yet impactful area for voice-enabled clinical tools.

2. Dataset

The Admedvoice dataset, a developed repository of Polish medical recordings, was used to fine-tune the speech synthesizer to produce medical-domain speech. Statistics from the portal's homepage, which contain the developed corpus of Polish medical speech, are shown in Figure 1. Recordings were obtained from medical personnel via the Admedvoice portal-embedded recorder, where physicians could record phrases from the provided text. Of the 107,031 recordings, 55,356 were female voices and 23,769 were male, with samples varying widely in duration and sampling rates, making it suitable for training medical-focused text-to-speech systems. The rest were multiple voice recordings made with an engineered, highly directional acoustic intensity probe in hospital operating rooms. The preprocessing involved extensive data cleaning to ensure high quality, eliminating corrupted or silent files, and normalizing audio samples. The aligner model and other processing tools extracted relevant audio features, filtered by length and quality, to create a dataset optimized for training. The summary of speech files per medical specialty is presented in Table 1.

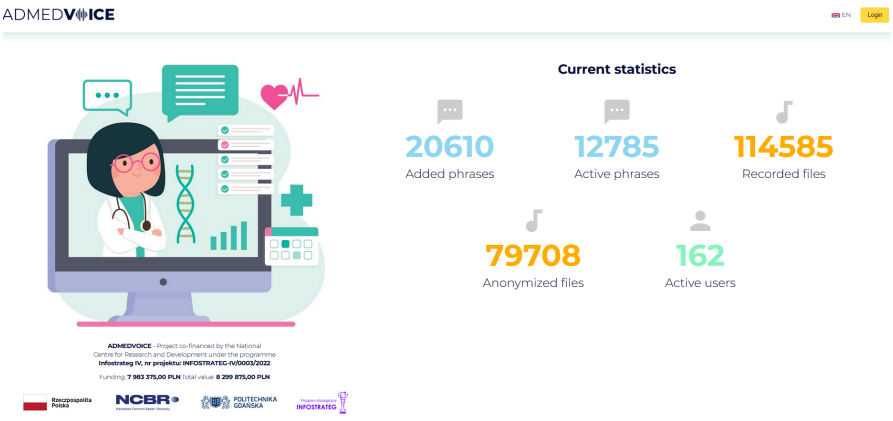

Fig. 1. The homepage of the Admedvoice repository that contains the dataset recorded in Polish. The page shows a numerical summary of the data.

2.1. Recording Preparation for Subjective and Objective Evaluation

For both subjective and objective assessments, recordings that comprised the test set had to be collected and processed. The phrases in the recordings were derived from medical research articles written in Polish. The speech excerpts were recorded in WAV format, lasting between 10 and 25 seconds, and came from the following sources:

- Human voice recordings made by the authors
- ElevenLabs synthesizer
- SpeechGen synthesizer
- Base ToucanTTS synthesizer
- ToucanTTS trained with Admedvoice recordings for 5 epochs
- ToucanTTS trained with Admedvoice recordings for 10 epochs
- ToucanTTS trained with Admedvoice recordings for 15 epochs

The phrases were chosen to represent a variety of medical terms commonly used in Polish medical contexts. Specifically, four phrases were selected for the MOS evaluation to ensure a balance between covering a broad range of medical terminology and preventing participant fatigue during the evaluation process. This selection allows for a comprehensive assessment without overwhelming the participants. The recordings underwent several processing steps. First, the number of audio channels was standardized from stereo (two channels) to mono (one channel). During preprocessing, approximately 10% of the recordings were removed due to corruption or excessive silence. The remaining recordings were then normalized to a peak amplitude of -3 dB to ensure consistent volume levels across all samples. Additionally, silent passages at the beginning and end of each recording were trimmed to focus on the relevant speech content. Finally, all recordings were resampled to a uniform sampling frequency of 44.1 kHz to eliminate any bias in the evaluation caused by differences in audio quality.

3. Methods

The following subchapters provide a brief overview of the methods used, starting with an exploration of speech synthesis and state-of-the-art tools in the field. A concise introduction to quality assessment metrics and a comparative analysis are presented in the subsequent sections. The chapter concludes with a discussion on objective and subjective assessments.

3.1. Speech Synthesis Methods

Speech synthesis can be divided into several methodologies with distinct characteristics and applications. Concatenative synthesis, one of the earliest techniques, involves assembling pre-

Table 1. Number of texts and recordings in the Admedvoice portal per medical specialty

	Texts	**Recordings**
1. Medical history	509	6325
2. Radiology - Thyroid	61	1758
2. Radiology - Breast	153	3641
2. Radiology - head, neck	457	5204
3. Oncology	822	5561
4. Pathomorphology	565	4254
5. Cardiology	463	5288
6. Course of surgical procedure	378	5207
7. Course of resuscitation operation	253	8673
8. Recommendations - hypertension	532	8676
9. Referral	465	9601
10. Prescription - Gastrointestinal	517	3868
10.B - Blood system	1008	5755
10.C - Cardiovascular	3178	6568
10.D - Dermatology	1713	14124
10.J - Anti-infective	799	6864
10.N - Central nervous system	832	5269

recorded speech segments to form continuous output. While this approach can yield high-quality, natural-sounding results, it is constrained by the availability of recorded units and often struggles with natural variations in prosody and intonation. Articulatory synthesis, on the contrary, attempts to model the physical processes of the human vocal tract to generate speech sounds. Although this approach provides flexibility in producing a range of phonetic sounds, it requires intensive computation and is less commonly used in practical applications. Statistical parametric synthesis provides an alternative approach, utilizing statistical models to generate speech waveforms from linguistic features. This method allows for greater control over prosody and intonation, but often requires a more natural sound associated with concatenative synthesis. However, profound learning-based synthesis has recently transformed the field, significantly improving naturalness and versatility. Models such as WaveNet, Tacotron, and FastSpeech employ neural networks to generate realistic audio waveforms or spectrograms directly from text, bypassing the need for recorded segments or complex physical simulations. These deep learning models represent the forefront of speech synthesis technology, setting new standards for quality and naturalness. While detailed architectural information for SpeechGen and ElevenLabs is not publicly available due to their proprietary nature, ToucanTTS is based on the FastSpeech2 architecture with enhancements such as language and speaker embeddings. These embeddings allow ToucanTTS to support multiple languages and generate speech in different voices. However, the reliance on FastSpeech2, which is optimized for general speech synthesis, may not fully capture the formal prosody and specialized vocabulary of medical speech. In contrast, SpeechGen and ElevenLabs likely incorporate advanced architectures and training techniques that enable them to produce more natural-sounding speech, particularly in specialized domains. Various studies have evaluated these systems against human voice standards [8, 9], highlighting their progress in achieving natural-sounding synthesized speech.

3.2. Quality Assessment Metrics

The research employed NISQA, a non-intrusive neural network-based metric, and subjective MOS to evaluate quality. NISQA provides objective ratings by analyzing distorted audio without reference files, producing scores from 1 (low quality) to 5 (high quality). At the same time, MOS reflects human perception by averaging listener ratings on a scale, typically from 1 to 5, or a more granular nine-point scale. These metrics provided a comprehensive view of the naturalness and overall quality of the synthesized speech.

3.3. Comparative Analysis

The comparative analysis aimed to evaluate the performance of different synthesizers based on objective and subjective assessments. The study included two state-of-the-art synthesizers (SpeechGen and ElevenLabs), as well as baseline and fine-tuned versions of ToucanTTS. The primary criterion for objective and subjective evaluations was voice naturalness rather than technical sound quality. This focus aligns with user preferences in real-world applications where naturalness significantly influences user experience.

4. Experiments and Results

ToucanTTS has been fine-tuned, and synthesis has been generated; more details can be found in section 4.1. After fine-tuning ToucanTTS, experiments for both scenarios —objective and subjective —were conducted to assess the quality of synthesis obtained from all tools. Objective and subjective approaches are described in sections 4.2 and 4.3.

4.1. ToucanTTS Fine-Tuning

The study's primary methodology focused on fine-tuning the ToucanTTS model using the Admed-voice dataset to produce more natural medical speech. The aligner model, which aligns text and speech features, was pre-trained to provide the text-to-speech model with structured input, enhancing synthesis accuracy. The fine-tuning was conducted for 5, 10, and 15 epochs due to computational constraints. As shown in Fig. 2, the training loss decreases steadily, suggesting that further improvements might be possible with additional epochs. However, to balance computational resources and time, we selected these intervals for evaluation. Future work could explore more extended training periods or alternative stopping criteria, such as monitoring validation loss, to optimize the fine-tuning process.

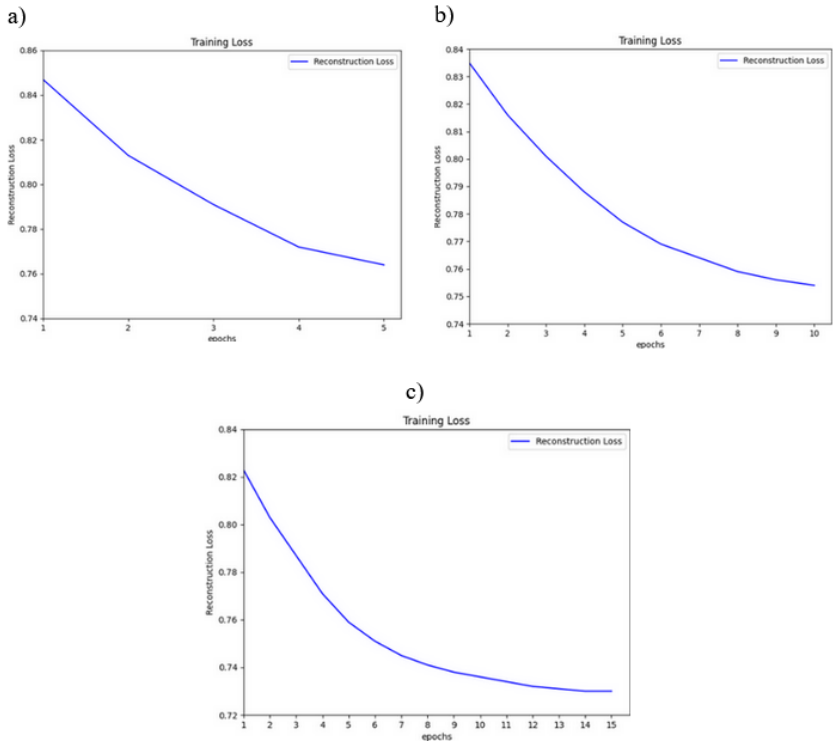

Fig. 2. Training loss for a) ToucanTTS - 5 epochs, b) ToucanTTS - 10 epochs, c) ToucanTTS - 15 epochs.

4.2. Objective Assessment Results

The synthesized speech from ElevenLabs and SpeechGen achieved statistically comparable NISQA scores to human speech, indicating a high level of naturalness. In contrast, all versions of the ToucanTTS model received significantly lower scores, suggesting they did not meet the same standards of naturalness as the leading synthesizers. The results are presented in Fig. 3.

4.3. Subjective Assessment Results

Participants provided ratings on a scale from 1 to 5, with 5 indicating the highest level of naturalness. The MOS results revealed that ElevenLabs and SpeechGen received average scores above 4.0, indicating that listeners perceived their outputs as highly natural. In contrast, all versions of ToucanTTS scored below 3.0 on average, highlighting a significant gap in perceived quality. The results are summarized in Figure 4, with detailed statistics provided in Table 2.

Table 2. Mean Opinion Scores (MOS), variance, and standard deviation for each synthesizer.

Synthesizer	Mean MOS	Variance	Standard Deviation
Human	4.08	0.50	0.71
ElevenLabs	4.19	0.52	0.72
SpeechGen	4.13	0.34	0.58
Base ToucanTTS	2.46	0.36	0.6
ToucanTTS (5 epochs)	2.27	0.27	0.52
ToucanTTS (10 epochs)	2.42	0.35	0.59
ToucanTTS (15 epochs)	2.34	0.37	0.61

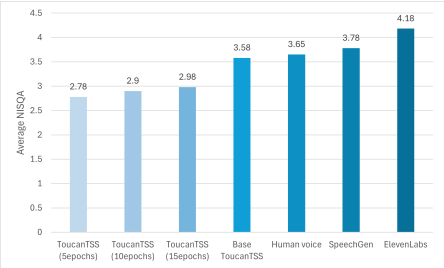

Fig. 3. Comparison of the objective quality of speech synthesizers.

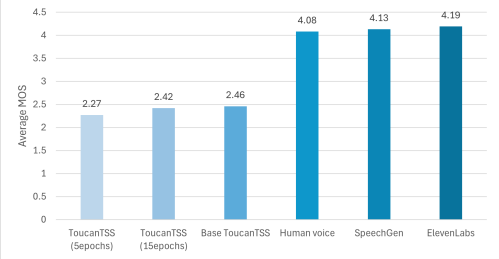

Fig. 4. Comparison of the subjective quality of speech synthesizers.

A total of 32 recordings were used for the subjective evaluation. This included four recordings each from human speech, ElevenLabs, SpeechGen, base ToucanTTS, and ToucanTTS fine-tuned for five epochs. Additionally, for ToucanTTS fine-tuned for 15 epochs, four recordings of the exact phrases and eight recordings of different phrases were included, totaling 32 recordings. This setup ensures comparability across synthesizers while also testing the robustness of the fine-tuned model on varied inputs. Participant feedback indicated that SpeechGen and ElevenLabs produced speech with better prosody, fewer artifacts, and more natural intonation compared to ToucanTTS. Specifically, listeners noted that the synthesized speech from these systems was smoother and more expressive, closely mimicking the cadence and emphasis of human speech. In contrast, ToucanTTS outputs were often described as robotic or monotonous, lacking the natural flow required for medical dialogues. The Mann-Whitney U test, a non-parametric test, $p < 0.05$ for all groups. This test allows for the comparison of two independent samples without assuming normality. The test was applied to the results presented in Figure 4, with a significance level set at alpha = 0.05. The results of the Mann-Whitney U tests for all pairwise comparisons are presented in Table 3.

Table 3. The statistical significance calculated for the pairs of compared sources.

Source 1	Source 2	p-values	Difference statistically significant?
ToucanTTS (5 epochs)	ToucanTTS (15 epochs)	0.013	Yes
baseline ToucanTTS	ToucanTTS (15 epochs)	0.374	No
ToucanTTS	Human voice	0	No
Human voice	SpeechGen	0.606	No
Human voice	ElevenLabs	0.049	Yes
SpeechGen	ElevenLabs	0.075	No

5. Conclusions

The study demonstrates that modern deep learning-based synthesizers, like ElevenLabs and SpeechGen, produce highly natural speech suitable for applications like clinical documentation and patient interactions. However, ToucanTTS requires extensive fine-tuning, high-quality data, and architectural improvements to achieve competitive performance in user perception and objective quality. Although it does not propose new models, this evaluation serves as a valuable benchmark, highlighting the current capabilities and limitations of TTS systems in underrepresented domains. It underscores the need for continued research in adapting TTS technologies to less-resourced languages, such as Polish, particularly in critical fields like healthcare.

5.1. Dataset Limitations and Augmentation

The Admedvoice dataset, while extensive with 107,031 recordings, consists primarily of medical personnel reading prompted texts. This may limit the prosodic variability and spontaneity typically found in natural medical speech, potentially reducing the effectiveness of fine-tuning for models like ToucanTTS. The lack of diverse speaking styles and emotional tones could hinder the model's ability to generalize to real-world medical dialogues. To address this, future work could employ data augmentation techniques such as pitch shifting, speed variation, or adding background noise to simulate more realistic speaking conditions. These methods can enhance the dataset's diversity and improve the performance of the fine-tuned model.

5.2. Challenges in Polish Medical Speech Synthesis

Polish medical speech presents unique challenges due to the language's complex phonetics, including consonant clusters and inflectional endings, which can be difficult for synthesis models to handle accurately. Additionally, medical terminology often requires a formal tone and precise pronunciation, which may not be adequately captured by models trained on general speech datasets. These factors likely contributed to ToucanTTS's underperformance compared to SpeechGen and ElevenLabs, which are proprietary models presumably trained on larger, more diverse datasets that include a wider range of speaking styles and domains. The specialized nature of medical speech necessitates models that can adapt to its unique characteristics, highlighting the need for domain-specific training data and architectural adjustments.

5.3. Practical Implications for Medical TTS Applications

The high naturalness scores (MOS > 4.0) achieved by SpeechGen and ElevenLabs suggest that these systems are suitable for immediate deployment in real-time clinical documentation or patient interaction systems for Polish-speaking doctors. Their ability to produce speech that closely rivals human naturalness can enhance user experience and acceptance in medical settings. In contrast, while ToucanTTS currently underperforms, its open-source nature makes it a candidate for further development, especially in low-resource settings where proprietary systems may not be accessible. With improvements in dataset quality and fine-tuning strategies, ToucanTTS could become a viable option for specialized applications. Given the estimated 165,000–170,000 Polish-speaking doctors globally, there is a significant market need for high-quality TTS systems in this domain. This study provides a foundation for selecting and improving TTS technologies to meet this demand.

5.4. Model Recommendations

Based on these findings, we recommend SpeechGen and ElevenLabs for immediate deployment in Polish medical TTS applications due to their superior naturalness. For research purposes, ToucanTTS shows potential but requires further improvements in dataset quality and fine-tuning

strategies to achieve competitive performance. Future research could also explore other open-source models such as VITS [10], which may offer better handling of Polish phonetics and prosody, potentially closing the performance gap with proprietary systems. Furthermore, while SpeechGen and ElevenLabs are proprietary systems, limiting their accessibility for research and development, ToucanTTS is open-source, providing opportunities for community-driven improvements and adaptations to specific domains, such as medical speech.

5.5. Summary of Key Findings

- SpeechGen and ElevenLabs achieve high naturalness (MOS > 4.0), making them suitable for medical applications.
- ToucanTTS underperforms but offers potential for improvement due to its open-source nature.
- Dataset limitations and architectural constraints likely contribute to ToucanTTS's lower performance.
- Future work should focus on larger, more diverse datasets and architectural enhancements to improve TTS systems for specialized domains like Polish medical speech.

6. Acknowledgements

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] Aaron Van Den Oord et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

[2] Yuxuan Wang et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[3] Yi Ren et al. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019. arXiv:1905.09263.

[4] Florian Lux et al. The ims toucan system for the blizzard challenge 2023. *arXiv preprint arXiv:2310.17499*, 2023.

[5] Speechgen homepage. Accessed: 2024-12-27.

[6] Elevenlabs homepage. Accessed: 2024-12-27.

[7] Admedvoice homepage. Accessed: 2025-01-07.

[8] Haibin Wu et al. Towards audio language modeling - an overview. *arXiv preprint arXiv:2402.13236*, 2024. 20 Feb 2024.

[9] Chen Chen et al. Hyporadise: An open baseline for generative speech recognition with large language models. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*, 2023.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021.