INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 4, PP. 361-366
Manuscript received February 20, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149553

Comparative analysis of natural and synthesized Polish speech

Michał Daniluk, and Agnieszka Paula Pietrzak

***Abstract*—In the evolving field of speech synthesis, not only intelligibility, but also naturalness remains an important factor. This paper presents a comparative analysis of natural versus synthesized Polish speech. Speech synthesizers: Ivona, Mekatron, Notevibes, and ttsmp3 were explored. Four methods for assessing synthesized speech quality and comparing it to natural speech were presented: the AB test, MOS, logatom articulation test, and MUSHRA. Sentence databases and a database of logatoms were generated for each synthesizer and recorded for natural speech. Results indicated natural speech was consistently better than synthesized speech. Among the synthesizers, Notevibes performed best in all comparisons, while Mekatron ranked lowest.**

***Keywords*—synthesized speech; AB test; MOS; MUSHRA; logatom articulation**

I. Introduction

SPEECH synthesis, the transformation of text into human-like speech, has been an area of exploration in the domain of linguistics and computational sciences for several decades [1]. As the modern way of living continues to evolve, communication technologies that concern synthesizing human voice are getting more popular. From the early days of simple computer-generated voices to the advanced voice assistants we use today, the development of speech synthesis shows the pursue to make machines sound more human. With the rise of applications like voice assistants, e-learning platforms, and assistive technologies, producing synthesized speech that is natural and mimics the human speech is getting more and more attention [2].

Traditionally, synthesized speech has often been perceived as robotic or unnatural compared to the dynamic qualities of natural human speech [3]. The nuances of human intonation, rhythm, and emotion have been challenging to replicate. However, advancements in deep learning and neural networks have revolutionized the field of text-to-speech (TTS) synthesis, offering potential enhancements in the naturalness and fluency of the generated speech [4].

In the context of the Polish language, with its distinct phonetic and grammatical features, assessing the quality and naturalness of synthesized speech can provide useful insight, in addition to objective measurements. Assessing the naturalness and quality of synthesized speech can be crucial to understand its effectiveness in emulating human speech. Listening test methods used for quality evaluation, such as AB, Mean Opinion Score (MOS), and Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests stand out for their reliability and comprehensiveness.

This research aims analyze four synthesizers: Notevibes, Mekatron, Ivona, and ttsmp3. Using standardized evaluation techniques: AB, MOS, and MUSHRA provides a consistent and objective lens for this assessment [5]. The findings will offer insights into the capabilities of current Polish speech synthesis systems.

II. Speech Synthesis

Speech synthesis, commonly known as text-to-speech (TTS), is converting written text into audible, human-like speech. Over time, several distinct paradigms have developed in its history, each characterized by its unique methods and features [1]. The first way of synthesizing speech was formant synthesis, manipulating sound waves based on phonetic rules, leading to efficient but often robotic sounding outputs [6].

The next approach was concatenative synthesis, which relied on reordering short segments of pre-recorded human speech to form new sentences. While many systems, including Ivona, have harnessed this method effectively, the challenge remained in ensuring seamless transitions between segments [7]. A shift then occurred towards statistical parametric synthesis, which utilized statistical models like Hidden Markov Models to produce speech, offering flexibility without the need for vast pre-recorded databases, albeit sometimes at the expense of naturalness [8].

The most transformative change in recent years has been the integration of deep learning and neural networks. Sequence-to-sequence models, such as Tacotron, have come to the forefront, capable of producing impressively natural speech, with Mekatron's adaptation of Tacotron2 being a notable example [9]. Further enhancing this are waveform generation models like WaveNet, which refine the synthesized audio to an even higher degree of realism [10]. The evolution of TTS showcases the field's dedication to bridging the gap between synthesized and genuine human speech.

III. Evaluated Speech Synthesizers

For this study, the focus was on synthesizers that support the Polish language and are universally accessible to every user. The synthesizers were selected based on the functionality of converting written text into speech, while those that offer voice recognition capabilities via recording devices were not considered. Four speech synthesizers were examined.

Michał Daniluk and Agnieszka Paula Pietrzak are with Warsaw University of Technology, Institute of Radiocommunication and Multimedia Technology, (e-mail: michal.daniluk2.stud@pw.edu.pl, agnieszka.pietrzak@pw.edu.pl).

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

A. Ivona

A widely recognized speech synthesis software offering Polish language support [7]. For the purpose of this study, two female voices (Ewa, Maja) and two male voices (Jacek, Jan) were used. Ivona provides adjustable volume and reading speed controls. Additionally, it offers the convenience of reading text files with options for sound output partitioning based on user preference. Files can be exported in .mp3, .ogg, or .wav format.

B. Notevibes

An online platform facilitating the creation of voice recordings in various languages and voice styles [11]. Generated files can be downloaded in .mp3 or .wav formats. This study utilized those voices available in lossless audio formats. Its user interface offers multiple settings, including the option to select silence duration after periods and commas in the analyzed text and reading style selection.

C. ttsmp3.com

A platform allowing users to convert entered text into speech, downloadable only in lossy .mp3 format [12]. It offers four distinct voices in the Polish language. One limitation is the maximum character count, which is set at 3000 for the text input field.

D. Mekatron

A Polish implementation of the TTS (text-to-speech) originally known as Tacotron2 [9]. It's an open-source project permitting users to interact with the trained voice model via Google Colab. In the utilized Jupyter environment, which supports both CPU and GPU processing, the source code is visible—an attribute not present in the other studied synthesizers. The decoder's maximum iteration number affects the length of the generated voice and can be increased for longer outputs.

IV. SOUND DATABASE

In order to carry out the study, it was necessary to create a database of sound samples, both from the use of synthesizers and those representing natural speech. The synthesized sound samples used were generated by four forementioned synthesizers. For each synthesizer default settings were used. Sound samples for natural speech were recorded in the anechoic chamber by male and female speakers. All samples were normalized to - 23 LUFS. Sound database consisted of sentences in polish and logatomes. The selection of the Polish sentences and logatomes ensured a comprehensive coverage of phonemes to evaluate the synthesisers' efficiency in producing varied speech sounds. To ensure consistency across all recordings, speakers were provided with specific instructions regarding tone, pace, and pronunciation.

A. Sentences

For the MUSHRA, AB test and MOS test, sentences created based on polish matrix test [13] were used. Matrix tests are a structured method for evaluating speech intelligibility. These tests typically consist of a matrix of words, wherein each row and column contains specific, predetermined words. When conducting the test, a random word from each column is selected to form a sentence. The matrix has columns for subjects, verbs, numbers, adjectives, and objects. A randomly generated sentence might be "John bought three red apples". The listener's task is to correctly identify each word in the sentence. This method ensures that the sentence structures remain consistent, while the content varies, providing a reliable means of assessing speech comprehension without the influence of predictability.

Based on the described matrix, 30 random sentences were created. For the AB test, 20 sentences were selected for each generator and for natural speech. In the MOS and MUSHRA test, only one sentence was used: „Tomasz nosi pięć dobrych piłek" ("Thomas carries five good balls").

B. Logatoms

Logatoms are words or syllables that have no linguistic meaning, so using them in test the test subject does not rely on language knowledge for speech comprehension. Logatoms for each synthesizer and natural speech were generated in four different variations [14]: vowel - consonant - vowel (VCV), consonant - vowel - consonant (CVC), consonant - consonant - vowel or vowel - consonant - consonant (CC) and several-syllable word logatoms.

V. METHODOLOGY

The study consisted of four listening tests. Tests were performed using headphone listening. The test stand consisted of a computer, a third-generation Focusrite Scarlett 2i2 audio interface and Beyerdynamic DT 990 Pro studio headphones. For the AB test, MOS and logatom articulation test, the study group was 30 people. The MUSHRA test was conducted with 50 subjects. Four testing methods were used to assess the quality of synthesized speech.

A. AB test

The AB test is a comparative methodology used to evaluate the nuances between two distinct audio signals, termed as 'A' and 'B'. Participants in this study are tasked with the objective of discerning and determining which of these two signals provides a superior auditory experience.

When focusing on the evaluation of synthesized speech, one of these signals will be artificial, synthesized speech generated with one of the four synthesizers. The other signal will be recorder natural human speech.

To avoid any biases that might arise due to the order of presentation, the audio samples are presented to participants in a randomized sequence. The aim of this process is to ascertain which of the speech synthesizers was more frequently preferred over natural speech by the study's participants.

B. MOS

Mean Opinion Score (MOS) was used to assess the quality of speech. Using this index, which is a rating scale from 1 to 5, participants rate the quality of the speech they hear based on their subjective feelings. The MOS index is defined in detail in the ITU-T P.800 standard [15] developed by the International Telecommunications Union. It contains principles and guidelines that are also commonly used in speech synthesizer evaluation studies.

MOS test was developed in Google Form, in which responses were collected on a scale of one to five to the question "How would you rate the wording of this sentence?". Subjects are asked to listen to a sentence generated by each of the four

synthesizers tested, as well as natural speech. Subsequent sounds are then rated on the naturalness of the voice on a scale of 1 to 5, where 1 indicates an "absolutely robotic" voice and 5 indicates an "absolutely natural" voice. After collecting the results from the study group, the arithmetic average of the ratings given by the listeners is calculated, which gives the MOS score. The obtained average values of each source are compared with each other. After comparison, it can be deduced which of the tested sources generates the most natural and closest sound to human speech. The higher the score, the sentence from a given source was judged to sound more natural.

C. MUSHRA

Multiple Stimuli with Hidden Reference and Anchor test (MUSHRA) is a method used to evaluate the quality of speech synthesizers, which was originally developed to assess the sound quality of audio transmission systems. The MUSHRA method is defined in detail in ITU-R BS.1534-3 [16]. Subjects are required to rate stimuli according to the Continuous Quality Scale (CQS). It consists of five equal intervals with adjectives denoting their quality [17].

Survey participants rate the quality of each sample on a point scale, usually on a scale of 1 to 100, where a value of 100 indicates the highest quality. This scale is used to evaluate the quality of each sample against a reference signal (reference). Among the signals to be evaluated is included a hidden reference, which is a sample of the same quality as the reference, but it is not revealed to the participants, so they can evaluate the quality of the samples without knowing which one is the reference. The hidden reference signal should be evaluated with a value close to the maximum. So-called anchors are also included between the signals being evaluated. Anchors are created by subjecting the reference signal to 3.5 kHz or 7 kHz low-pass filtering. This produces two signals containing an incomplete frequency band compared to the reference. Due to this filtering, they are expected to be evaluated close to the minimum value. If these conditions are not met, the results are considered unreliable.

The MUSHRA method has the advantage that the samples presented can be listened to in any order, and the scales for evaluating the signals are displayed simultaneously so that the subject is able to make a direct comparison between them. Existing software [18] was used to conduct the MUSHRA test.

D. Logatom articulation test

Examining logatom articulation can be done by a presentation of a set of logatomes to test respondents, who indicate, what logatome they heard. The test was developed in Matlab and Matlab App Designer. The software was programmed to randomly select and play a logatome from a predefined set [19], with each test participant responding by typing in the logatome they believe they heard. Based on the responses, the logatom articulation score was calculated, as a percentage of correct recognition of the presented logatoms [20].

VI. RESULTS

Four test were conducted to compare synthesizers to natural speech and to assess the synthesizers quality. The results are presented separately for every test type. Each section shows the findings and conclusions for the given testing method.

A. AB test

The vast majority of study participants showed a preference for natural speech over the other sources. In many instances, all 20 participants indicated that the natural speech sounded better. Only in a few cases was natural speech rated lower, but it still dominated. For 2 listeners, the synthesized speech scored a maximum of 6 out of a possible 20 points, indicating that even though some participants might have favored its sound, it wasn't as popular as natural speech.

The Ivona speech synthesizer secured very low scores, a maximum of 2 out of 20. This suggests it was rarely favored by participants. Similarly, the Mekatron synthesizer only garnered 2 points in the entire test, pointing to its infrequent selection as sounding better. The Notevibes and ttsmp3 platforms had somewhat similar scoring distributions, occasionally achieving 2 or 3 points. However, just like other synthesizers, they were rarely preferred compared to natural speech.

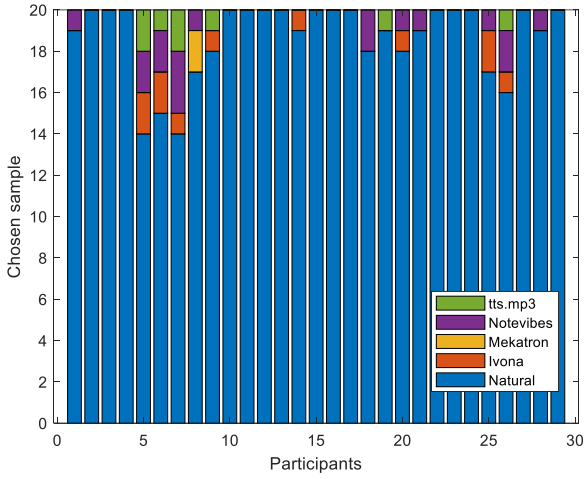

Fig. 1. AB test results - distribution of chosen sound samples by 30 participants, comparing natural speech to four different synthesizers.

Participants showed a strong preference for natural speech over the four speech synthesizers. Among the synthesizers, none stood out as a consistent favourite, but Notevibes and ttsmp3 appeared to be occasionally preferred by a few participants.

B. MOS

Obtained results are presented as the box-and-whiskers diagram (Fig. 2). Natural speech, consistently achieved higher scores, mostly obtaining 5s and occasionally 4s. This underscores listeners' preference for the natural human voice, which often acts as a standard in such evaluations.

As for the synthesizers, listeners predominantly found Ivona's quality to range between poor to average, as most of its scores hovered around 1, 2, or 3. Mekatron was given scores of 1 and 2, indicating a perception of its quality as being between poor and fair, although there were a few instances where it received a 3. Notevibes exhibited a more diverse set of scores, fluctuating from 1 to 5, revealing mixed feedback from listeners about its quality. Lastly, tts.mp3's scores mainly oscillated between 1 and 3, denoting a perceived quality ranging from poor to average.

While natural speech consistently surpassed the other voice sources in quality, synthetic voices displayed a broader spectrum of listener perceptions. Among these, Notevibes evoked mixed reactions, with some deeming its quality as excellent. In contrast, Ivona and Mekatron were generally perceived to be of inferior quality.

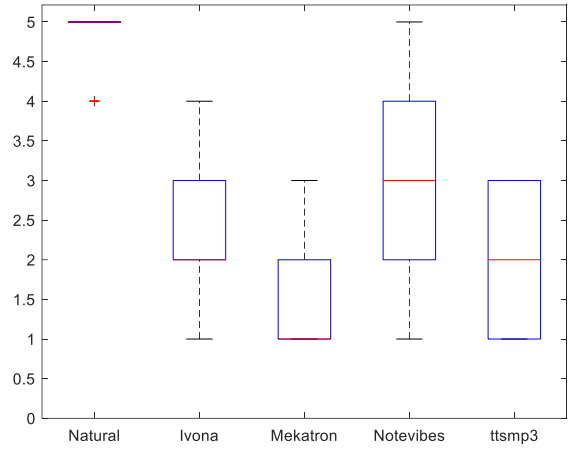

Fig. 2. Mean Opinion Score (MOS) results - box plot representation of user ratings for natural speech and four speech synthesizers, highlighting the median, quartiles, and outliers.

C. MUSHRA

Out of the 50 individuals tested, only 12 of them met the criteria for a credible listener (a score of >90 for the reference and a score of <10 for the anchors). The rest were assessed as non-credible.

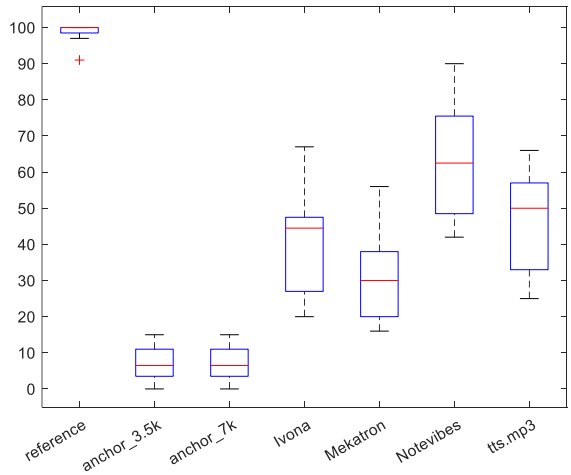

Fig. 3. MUSHRA test results - box plot comparison of perceived quality for reference (natural speech), anchors and four speech synthesizers, indicating median, quartiles, and outliers.

In the obtained results (Fig. 3), the reference, representing natural speech, consistently scored near 100, showcasing its superior quality. The Mekatron synthesizer scores displayed notable variability, predominantly skewing toward the lower end, with values ranging between 16 and 56. In contrast, tts.mp3 demonstrated a mid-tier performance with scores oscillating between 25 and 66. Notevibes often scored high, with values hovering between 45 and 90, indicating its quality was often perceived as nearly comparable with the reference. Meanwhile, Ivona exhibited a broader range from 20 to 67, positioning it in the mid to high-quality spectrum but lacking the consistency of the reference or Notevibes.

In essence, while the reference, so natural speech unsurprisingly scored best, Notevibes was scored best among the evaluated synthesizers, and Mekatron was ranked worst in perceived quality.

D. Logatom articulation test

The test assessed how different types of speech were understood in terms of recognizing logatoms. Obtained results on the percentage of correct recognitions are presented as box plots for natural speech and four assessed synthesizers (Fig. 4).

Natural speech consistently outperforms the speech synthesizers in logatom articulation, often nearing or even reaching 100%. This suggests that human speech is much clearer in recognizing logatoms. Among the synthesizers, Ivona displays a broad spectrum of results, sometimes nearing the performance of natural speech, but with noticeable variability. Mekatron tends to score lower, frequently falling below the 50% mark, indicating a greater challenge in recognizing logatoms. Notevibes offers a more stable performance, predominantly around the 70% range, while ttsmp3's performance parallels Notevibes but with some inconsistency, occasionally reaching up to 80% but also dipping at times. In summary, while natural speech remains the benchmark for clarity, there's a notable variation in performance among the speech synthesizers.

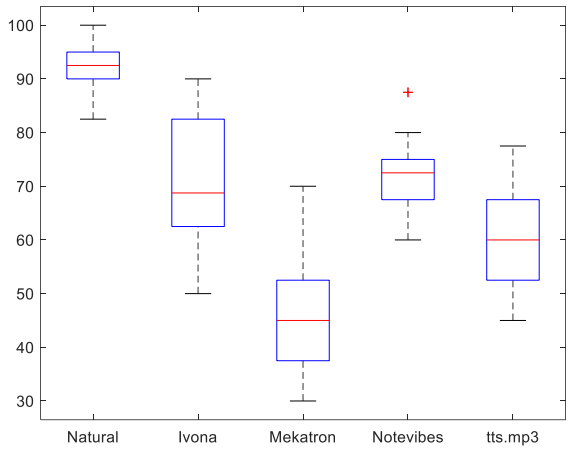

Fig.4. Logatom articulation test results - box plot illustrating the logatom articulation based on percentage of correct recognitions for natural speech and four speech synthesizers.

VII. Discussion

A key insight is the persistent preference for natural speech. The preference may be based on the authenticity of a natural voice, its clarity, tonal variations, and emotional content. The AB tests confirmed this trend, with 93.5% of trials showing a preference for natural speech compared to its synthesized versions.

Further analyzing the AB test, in only about 6.5% of all trials, the participants felt that synthesized speech sounded better. Among these rare results, Notevibes was the preferred choice,

chosen in almost 12% of comparisons, considering 5 trials in each study where it appeared. Mekatron achieved the lowest score, at just under 1.4%.

The gathered MOS ratings suggest that human speech is rated as the most natual. A near-maximum score of 4.83 validates the integrity of the conducted study. Among the speech generators, Mekatron was characterized by the most artificial sound. Ivona and ttsmp3 samples were rated at 2.4 and 2, respectively. Notevibes achieved the highest score among synthesized speech representations, approximately scoring 3.

In the MUSHRA study, listeners rated the reference signal, which is natural speech, as the highest. Among the synthesizers, Notevibes emerged as the top performer, while Mekatron secured the lowest score among the generators. Anchors received the poorest ratings. Notevibes achieved a score of 62.75, while Mekatron scored just under 31 on a 0-100 scale. This is a significant difference, considering that both samples were generated as lossless .wav files. Ttsmp3, on the other hand, which uniquely generated sound in the .mp3 format at a bit rate of 62kbps, achieved the second-best score among synthesized speech.

Only 12 MUSHRA results met the criteria of having a sufficiently high reference rating and a sufficiently low rating for limited-spectrum signals. 38 unreliable results were discarded. A significant number of unmet responses might be attributed to listeners basing their evaluation not just on sound quality but also on subjective preferences regarding sound. The challenge was further compounded by the fact that each presented sample represented a distinct voice. Listeners, when switching between signals, primarily focused on the most evident differences between them.

The averaged results of the logatom articulation study indicate that the syllables and words easiest to recognize originated from the natural speech database. Almost 94% of logatoms were correctly identified and recorded within this category. Aggregating and averaging the correct responses from all four generators revealed an approximate logatom articulation of 60%. When examining synthetic speech results separately, Ivona and Notevibes both stood out with higher scores (71% and 72%, respectively). Ttsmp3 samples showed a logatom articulation of 60%, while Mekatron lagged behind with a score of 45%.

Globally, the superiority of natural speech over synthesized speech was evident in every examined aspect. Comparing the tested synthesizers, Notevibes consistently showed the highest scores in terms of quality, logatom articulation, and naturalness of sound. In the AB test, when a listener decided that the synthesized signal sounded better, Notevibes was most often the chosen preference.

CONCLUSION

In this study an analysis was undertaken to assess the perceived quality of natural speech versus various synthesized speech samples. Series of listening test were conducted, measuring listeners preference, naturalness of presented sound and their overall quality. Several findings were made concerning the performance and perception of natural speech versus various speech synthesizers.

Across all tests, natural speech consistently emerged as the preferred and most effective form of communication among listeners. Its clarity and familiarity ensured that it was often recognized and rated higher than the synthesized speech.

While some synthesizers, like Notevibes, occasionally matched the performance of natural speech in certain tests, they still fell short in direct comparisons. This indicates that while technology has come a long way in replicating human speech, there's still a room for improvement.

Among the synthesizers, performance was varied. Notevibes often stood out as the most competitive, sometimes nearing the quality of natural speech. In contrast, Ivona and Mekatron was generally perceived by the listeners as of inferior quality. The tts.mp3 synthesizer was assessed as middle-quality, with good but inconsistent results.

The logatom articulation test provided results which showed that natural speech was undeniably superior, with synthesized voices, especially Ivona and Notevibes, displaying decent performance but also revealing areas for improvement.

It's worth noting that a majority of the study participants leaned towards natural speech when given a choice. However, the credibility of listeners played a role, especially in the MUSHRA test where a significant proportion didn't meet the criteria.

In conclusion, while speech synthesizers are advancing and offer a promising alternative to natural speech, they are yet to achieve the clarity, consistency, and overall preference that comes with the human voice. Continued innovations in this field might narrow the gap, but for the time being, natural speech continues to stand unmatched in its effectiveness and perception. Future research might give more insight into improving synthesizer technology or understanding the nuances of why listeners prefer the authentic human voice.

REFERENCES

[1] D. H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., vol. 82, no. 3, pp. 737-793, 1987. https://doi.org/10.1121/1.395275

[2] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, "A review of deep learning based speech synthesis," Applied Sciences, vol. 9, no. 19, p. 4050, 2019. https://doi.org/10.3390/app9194050

[3] T. Dutoit, "High-quality text-to-speech synthesis: An overview," J. Electrical & Electronics Engg., vol. 17, no. 1-2, pp. 25-33, 1997.

[4] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: A review," Artificial Intelligence Review, vol. 56, no. 7, pp. 5837-5880, 2023. https://doi.org/10.1007/s10462-022-10315-0

[5] ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunication Union, 2015.

[6] J. L. Flanagan, "Speech analysis, synthesis, and perception," Springer, 1972.

[7] M. Kaszczuk and L. Osowski, "Evaluating IVONA Speech Synthesis System for Blizzard Challenge 2006," Blizzard Workshop, Pittsburgh, PA, 2006.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, 2009.

[9] Tacotron2, “Tacotron 2 synthesis”, Google Colab, 2023. [Online]. Available: https://colab.research.google.com/drive/1gsPMm4mBD71WcTftEffMs3-N89HlD1ju

[10] A. van den Oord et al., "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[11] Notevibes, "Polish text-to-speech," Notevibes, 2023. [Online]. Available: https://notevibes.com/polish-text-to-speech/

[12] TTSMP3, "ttsmp3 API Documentation," TTSMP3, 2023. [Online]. Available: https://ttsmp3.com/apidoc.php

[13] E. Ozimek, A. Warzybok, and D. Kutzner, "Polish sentence matrix test for speech intelligibility measurement in noise," International Journal of Audiology, vol. 49, no. 6, pp. 444-454, 2010. https://doi.org/10.3109/14992021003681030

[14] J. Rafałko, "Algorytmy automatyzacji tworzenia baz jednostek akustycznych w syntezie mowy polskiej," Institute of Systems Research of the Polish Academy of Sciences, 2014.

[15] International Telecommunication Union, "Recommendation I.T.U.T. P. 800: Methods for subjective determination of transmission quality," Geneva, 1996.

[16] International Telecommunications Union, "Recommendation I.T.U.R. 1534-1: Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," Geneva, Switzerland, 2001.

[17] International Telecommunication Union Radiocommunication Assembly, "Method for the subjective assessment of intermediate quality level of audio systems," Series B, 2014.

[18] W. Bartosik, "Projekt i realizacja aplikacji webowej do tworzenia i przeprowadzania testów słuchowych MUSHRA," Institute of Radioelectronics and Multimedia Technology, Warsaw University of Technology, 2020.

[19] S. Brachmański, "Test material used to assess speech quality in Poland," in Acoustics, Acoustoelectronics and Electrical Engineering, F. Witos, Ed., Gliwice, 2021, pp. 65-79.

[20] S. Brachmański, "Selected Issues of Speech Signal Transmission Quality Assessment [Wybrane zagadnienia oceny jakości transmisji sygnału mowy]," Wrocław, Poland: Oficyna Wydawnicza Politechniki Wrocławskiej, 2015.