INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1-6
Manuscript received May 19, 2025; revised July 2025. doi: 10.24425/ijet.2025.153612

Quality assessment of synthetic speech

Stefan Brachmański, Maurycy Kin, and Piotr Kozłowski

***Abstract*—This paper presents the results of a subjective study of the quality assessment of several selected speech synthesizers. The subjects of the study were logatom intelligibility and overall speech signal quality evaluation. Synthesizers generating both male and female voices were used for the study. An attempt was also made to apply objective quality assessment methods used to test the quality of transmission in telecommunications channels. The results of these attempts, however, showed the impossibility of using the PESQ method to assess the quality of synthetic speech, mainly due to the lack of temporal synchronization between the test signal and the reference signal.**

***Keywords*—speech quality; speech synthesis; logatom intelligibility**

I. Introduction

THE first attempts to transform human speech by machine were made in the second half of the 18th century. In 1773, Kratzenstein, a professor of philosophy from Copenhagen, successfully generated vowels using resonance tubes connected to an organ. At the same time, Wolfgang von Kempelen constructed a machine that generated whole words and even short sentences in Latin, French, and Italian. Over the centuries, there have been many attempts to construct speaking machines, but it was not until the advancement of electrical engineering in the early 20th century that the synthesis of speech sounds by electrical means was possible. The first device of this type was the "Voder" (Voice Demonstrator), constructed by Homer Dudley and presented in New York in 1939 [1,2]. The technology of that time based on hardware solutions was continued for many years. Further development of speech synthesis took place in the 70s with the development of computer technology. [3, 4, 5].

Modern synthesizers are implemented as software solutions. At the current state of technology, the limits of achievable intelligibility and naturalness of speech synthesis are no longer set by technological factors, but by our limited knowledge of acoustics and speech perception. Modern synthesizers also use artificial intelligence (AI) technologies, enabling the conversion of any text into speech (Text-To-Speech – TTS) [6,7,8]. In information and dialogue systems, the role of a human is taken over by a virtual operator creating their statements using artificial intelligence [9,10,11]. Recent advances not only make it possible to produce human speech, but also to determine the gender and age of the person uttering the words. For example, it is possible to determine whether the produced speech corresponds to a 60-year-old man, a 30-year-old woman, or a 12-year-old child.

The aspects of speech quality is connected usually to the transmission rate in telecommunication channels, or digital broadcasting [3]. The characterizations are broadcast quality at bit rates exceeding 64 kb/s, communications quality at 6 to 12 kb/s and synthetic quality at bit rates of 6 kb/s, or lower. Quality of synthetic speech is substantially less natural than broadcast or communications quality but has essential intelligibility.

From a quality viewpoint, a TTS system is more complex than a speech coder in which the human input speech is encoded at various bit rates, transmitted and stored, and then decoded and delivered to the receiver with some degradation due to information loss. In TTS system, the speech is generated by recomposing the words and sentences from a finite set of synthesis blocks as phonemes, diphones and other speech elements. It may improve the naturalness of the speech [12].

Various speech quality factors must be satisfactory for good communication, however, in the higher quality range, naturalness of speech is required while intelligibility of speech is the most important factor in the lower quality range [13]. A very important elements of synthetic speech are its overall quality and intelligibility. Speech intelligibility is one of the basic quality parameters of speech signal transmission in both analog and digital telecommunications chains, as well as in auditoriums and the sound systems used in them, verbal alerting systems, synthesizers, and in the selection of hearing aids. Speech intelligibility is not the same parameter as overall speech quality. For example, the speech signal emitted in warning systems may not sound pleasant, however, warning messages must be conveyed in an effective, understandable manner. Assessment of the quality and intelligibility of synthetic speech can be performed by various subjective [14,15,16,17] or objective [18,19,20] methods. In the present study, the evaluation of the quality of synthetic speech was performed on both intelligibility and qualitative criteria using logatom intelligibility [14,15] and Absolute Category Rating [21].

On the other hand, the use of objective measurement methods for assessing the quality of synthetic speech, such as PESQ (Perceptual Evaluation Of Speech Quality) [22] or POLQA (Perceptual Objective Listening Quality Analysis) [23], would simplify the quality control procedure, through the possibility of automating measurements and random quality checks. However, the principle of these methods is based on comparing the tested signal with a standard, and in the case of synthetic speech - there is no such standard.

The purpose of the presented research was to check the quality of speech generated by the most popular speech synthesizers by the means of subjective assessment. An additional aim was to investigate the potential of employing objective quality assessment methods to evaluate synthetic speech.

First Author, the second and the third Authors are with Wroclaw University of Science and Technology, Poland (e-mail: stefan.brachmanski@pwr.edu.pl, maurycy.kin@pwr.edu.pl, piotr.kozlowski@pwr.edu.pl).

II. RESEARCH METHOD

A. Measurement of speech intelligibility by the logatom method.

One of the basic quality parameters of a speech signal is intelligibility. Measurement of speech intelligibility can be based on sentence lists (sentence intelligibility) or logatom lists. In the present study, quality assessment in terms of intelligibility of synthetic speech was based on logatom intelligibility. Measurement of logatom intelligibility can be carried out in the traditional version or with an alternative choice.

In the traditional method, the listener writes down the received logatoms in orthographic form, and then the expert checks the correctness of the received logatoms according to the phonetic rules. The final step is to calculate the average logatom intelligibility as the ratio of correctly received logatoms to all logatoms generated according to equation (1) [14].

$$W_L = \frac{1}{N \cdot K} \sum_{n=1}^{N} \sum_{k=1}^{K} W_{n,k} \quad [\%] \qquad (1)$$

where: N – number of listeners, K – number of presented lists, $W_{n,k}$ – logatom intelligibility of k-numbered test list obtained by the n-th listener, while:

$$W_{n,k} = \frac{P_{n,k}}{T_K} \cdot 100 \quad [\%] \qquad (2)$$

The logatom intelligibility measurement method with selection is an automated version of the traditional method. The measurement process is controlled from a computer with a sound board. Monophonic sound files, such as logatoms generated by a synthesizer, are placed on the hard disk.

TABLE I
SPEECH QUALITY CLASSES FOR THE TRADITIONAL LOGATOM INTELLIGIBILITY MEASUREMENT METHOD

Class	Characteristics of the quality class	Logatom intelligibility
1	Comprehension without the slightest strain of attention, Signal without noticeable contamination	> 75 %
2	Comprehension without difficulty, Subjectively noticeable signal contamination	60 - 75 %
3	Comprehension with focused attention, Without repetition or questioning	48- 60 %
4	Comprehension with high attention span, with repetition and questioning	25 – 48 %
5	Inability to fully understand, breaking the connection	<25 %

The program controlling the measurements retrieves the test signal (logatom) from the audio database and presents it to the listener. At the same time, seven alternative logatoms are presented textually on the monitor screen [14]. Among these seven logatoms, the presented logatom is placed in a random position. The listener's task is to choose the correct - in his opinion - answer. After the listener selects the test item number, the program compares the versions of the logatom selected with the logatom given. When the result of the comparison proves positive, the value of the variable storing information about the correctly received logatoms is increased. When the measurements are completed, the logatom intelligibility result is displayed on the monitor screen. On the base of the logatom intelligibilty obtained, speech quality classes can be determined [14], what is presented in Table I.

B. Measurement of speech quality by the method of Absolute Category Rating (ACR).

The well-known Absolute Category Rating (ACR) method recommended by the International Telecommunication Union (ITU) [21] can be used to test the quality of synthetic speech. This method uses test lists composed of simple, short, semantically unrelated sentences. The list is divided into groups of five short (2-3 s) sentences. There is a silence interval of 8-10s between each group. After listening to the group of sentences, during this pause, the listeners give a quality rating on a five-point scale. On this scale, a rating of 5 means very good quality, 4 - good, 3 - sufficient, 2 - pure, 1 - insufficient.

It is also possible to perform experiments using the MUSHRA method [24]; however, recent literature reports suggest a high convergence of subjective assessment results between the two methods [16].

The average (final) score is calculated for each speech synthesizer tested as the result of averaging by listeners. The score is given as the Mean Opinion Score (MOS). One advantage of the MOS scale is that different impairment factors can be evaluated simultaneously and the listener's opinion can be assessed directly. One must remember, however, that the test conditions for MOS methods are very specific and every exception from the recommendations may affect the results.

C. Test material.

The test material used in measuring speech quality and intelligibility included phonetically balanced sentence and logatom lists [25]. The condition of phonetic balance means that the percentage of individual phonemes in the test list should coincide with the frequency of these phonemes in Polish speech.

The sentences were recorded using 10 different synthesizers with which the utterances of 5 male and 5 female voices were generated. A total of 500 samples were generated, i.e. 50 sentences for each synthesizer.

During the experiment, the listeners' task was to listen to short sentences recorded with different synthesizers and give their opinion on the speech they listened to, on a five-point MOS scale or write down the logatom they perceived.

In the comprehensibility-based part, logatom lists were used, while in the part based on the quality criterion, the test material consisted of simple, easy-to-understand short sentences. The sentences were randomly arranged in a random order so that there was no direct connection to the subsequent sentences. Sentences ranged in length from 2 to 3 seconds. The database created had a total of 500 sentences, with 250 sentences generated by the synthesizer with a female voice and 250 with a male voice.

Ten types of software synthesizers were used to study the quality of synthetic speech:

1. Synthesizer WP (male voice)
2. Realspeak (female voice - Agata)
3. Syntalk (male voice)
4. eSpeak (male voice)
5. eSpeak (female voice)

6. mySimpleSynth (female voice)
7. Acapela (female voice - Ania)
8. Expressivo (male voice - Jacek)
9. Expressivo (female voice - Ewa)
10. Dant Free (male voice)

The acoustic signal was recorded in monophonic-channel in PCM format with a sampling rate of 16,000 samples/s and a resolution of 16-bit.

D. *Listening team and measurement technology.*

A listening group consisting of people experienced in subjective tests took part in the measurements [26,27]. The same people also participated in a study of the quality of radio broadcasts transmitted on the single-frequency DAB+ digital radio network [26]. Thus, the listening group consisted of people between the ages of 18 and 30. The group size was 30 people. All subjects were briefed on the principles of measurement. Listening to sentence lists was performed using headphones Each measurement consisted of one list, which was divided into 5 groups of 5 sentences in each. Each group was phonetically balanced. Examples of the 2 groups of Polish sentences list used for synthetic speech generation are presented in the Table II.

TABLE II
EXCERPT FROM A SENTENCE LIST

Group 1	Group 2
Ojciec podniósł się od stołu.	Ludzie udzielają pomocy rannym.
Znalazłem coś dla siebie.	Były to dla mnie najmilsze chwile.
Płomienie oświetlały żołnierzy.	Wojtkowi podobał się kolor.
Przyjaciółki powinny sobie pomagać.	Zła sytuacja wymaga zmian.
Wkrótce usłyszałem helikopter.	Widzę że jesteś w świetnej formie.

During the test, listeners were tasked with evaluating short sentences presented by the program at equal intervals. After each group of sentences, the test subject had eight seconds to evaluate the signal they heard. The samples were rated on a five-point quality scale. Each listener listened to and rated the letters for 10 different synthesizers.

In the logatom intelligibility measurements, the listening group consisted of subjects aged 20 to 25 with normal hearing. The group size was also 30 subjects [26]. All subjects were familiarized with the principles of measurement. All listening was performed using stereo headphones. Each measurement consisted of one list, which included 100 logatoms.

During the logatom intelligibility test, the listeners' task was to evaluate logatoms broadcast by the program at equal intervals. The exposure time of the logatoms was set at 4 seconds. During this time, the listener had time to decide which of the exposed seven test logatoms was the presented one. Each person listened to and evaluated 10 sets of logatoms with the same 10 different synthesizers.

III. RESULTS

A. *Assessment of the quality of synthetic speech by measuring logatom intelligibility*

The results of synthetic intelligibility averaged for the entire listening group for the synthesizers tested are shown in Figure 1. In addition, the standard deviation values of the results obtained are shown in Table III. For statistical verification, the homogeneity of the variances of the responses of individual listeners was tested using the Bartlett test based on the $\chi 2$ statistics. The test was performed at a significance level of α = 0.05. The test showed that the variances of the listeners' responses were homogeneous ($p > 0.05$), except for the synthesizer eSpeak (male voice) ($p = 0{,}0016$).

The obtained logatom intelligibility percentages range from 37.9 to 83.4%. The highest logatom intelligibility value was obtained for the Acapela synthesizer generating female voice (83.4%), that situates it in the highest, first class of devices. In contrast, the lowest intelligibility value was obtained for the WP synthesizer generating a male voice (37.9%). In this case, the device belongs to the fourth quality class. The rest of the tested devices have logatom intelligibility values in the range of 48% - 78%, i.e. in terms of logatom intelligibility they were rated as good and very good, according to the criteria [14,15,28].

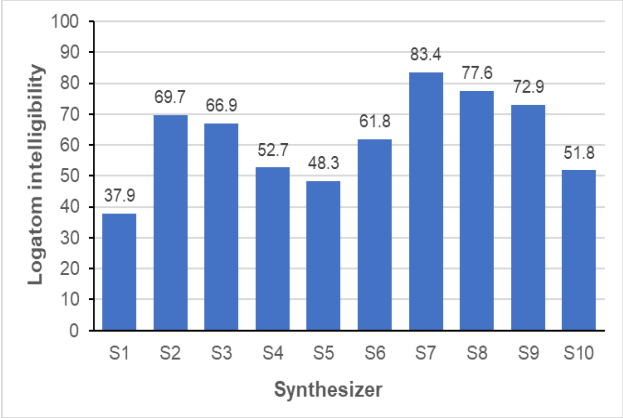

Fig. 1. Logatom intelligibility for tested synthesizers

Also noteworthy is the significant value of the standard deviation of the results for the WP synthesizer, almost 2 times higher than for the other synthesizers investigated. Therefore, it was decided to examine the conformity of the distribution of scores for each synthesizer to a normal distribution. For this purpose, the Kolmogorov-Smirnov test of the distribution conformity of scores with the normal distribution was used, at a significance level of α = 0.05. The results of the analysis indicated that for eight of the devices studied, the distribution of logatom intelligibility ratings followed a normal distribution. The probability p-values ranged from 0.0639 to 0.0160. It should be noted that these results included synthesizers for which the highest logatom intelligibility rating was obtained. For the other two synthesizers: WP and eSpeak (male voice), on the other hand, the distributions of the expressivity score were shown to follow a Poisson distribution (p = 0.032 for WP and p = 0.019 for eSpeak - male voice, respectively). This means that the results are distributed asymmetrically, with a tendency similar to a binomial distribution, where listeners present very precise criteria, on a 0 - 1 response basis.

B. *Quality assessment of synthetic speech with ACR method*

Figure 2 shows the MOS (Mean Opinion Score) scores averaged over the whole listening group, while Table III shows the standard deviation values of the scores of all listeners.

After rejecting the results of the only one subject whose responses did not meet the Chauvenet criterion [12], the

homogeneity of the variances of the responses of individual listeners was tested using the Bartlett test based on the $\chi 2$ statistics. The test was performed at a significance level of $\alpha = 0.05$. It turned out that the variances of the listeners' responses were homogeneous for all the synthesizers tested ($p = 0.0829$).

.

TABLE III
THE STANDARD DEVIATION VALUES OF LOGATOM INTELLIGIBILITY AND QUALITY ASSESSMENT OF SUBJECTIVE MEASUREMENTS OF SYNTHETIC SPEECH GENERATED BY THE SYNTHESIZERS

Synthesizer	Standard deviation of logatom intellibility [%]	Standard deviation of quality assessment [MOS]
S1	8,5	0,22
S2	4,5	0,38
S3	4,6	0,18
S4	4,2	0,16
S5	5,0	0,02
S6	3,7	0,34
S7	4,4	0,12
S8	3,8	0,22
S9	4,4	0,21
S10	4,1	0,22

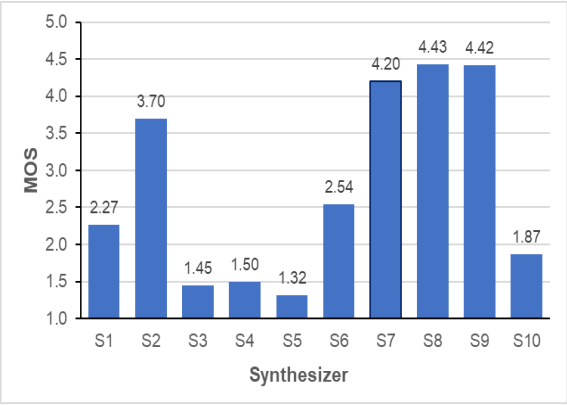

Fig. 2. Quality assessment as the listener-averaged MOS values for the synthesizers tested in the experiment.

Analyzing the results obtained for the test sentences by the subjective method of evaluating the quality of synthetic speech, it was found that the best synthesizer, featured the highest average evaluation score, is Expressivo for both female and male voice. The average of all listeners for the female voice is 4.42, while for the male voice it is 4.43. According to the ITU-T P.800 recommendation, the speech signal transmission quality can be considered as good at MOS - 4.0. Therefore, it should be concluded that in all these cases, the quality can be said to be better than good (MOS > 4.0). Also, the quality of the speech signal generated by the Realspeak synthesizer, producing female voice, was found to be only slightly worse than good (MOS = 3.7). The worst listening quality was achieved by the eSpeak synthesizer (female voice), which was only 1.32 of MOS. However, it was in its case that the standard deviation was the smallest, i.e. the average precision of quality assessment for all listeners was the highest, which shows the high degree of univocality of the evaluation criterion adopted by listeners. The sound quality of speech generated by the other synthesizers was rated much lower - (MOS < 2.54), which means that they can be used in special applications where sound quality is not the key.

It should be noted that the WP synthesizer was rated as the worst in terms of the logatom intelligibility index. In view of the above, for the remaining eight speech synthesis devices for which the variances were found to be homogeneous, an ANOVA test based on the *F*-Snedecor statistic was performed to determine the statistical significance of the difference in logatom intelligibility. As a result of this analysis, it was found that the differences in the values of the subjective assessment of logatom intelligibility were statistically significant ($p = 0.00389$).

C Attempts to implement objective evaluation methods to study the quality of synthetic speech.

The QE-ARM method developed at Wroclaw University of Science and Technology [25], was used to evaluate synthetic speech. The main aim of application of the QE-ARM method is to recognize samples of speech signals subject to transmission in digital telecommunications chains. It compares a reference sample with a sample at the output of the channel, similar to the PESQ or POLQA method. The algorithms used in this method are based on automatic speech recognition procedures. The most important parameters for configuring the algorithms are the number of LPC parameters, for a prediction order of 8 to 32, the choice of metric, the number of time classes, or the parameterization method (FFT, LPC and Cepstrum). The same logatom lists were used for testing this method as in the subjective measurements, while excerpts from other logatom lists, prepared for digital radio transmission quality studies [29], were used as reference signals. After testing several logatom lists, it turned out, unfortunately, that the objective method used to assess speech quality is not suitable in this case. The fundamental reason for this is the lack of correspondence between the representation of the two signals: the reference and the tested one in the time domain. The best result obtained in the study was 13% for the S7 synthesizer, for which the subjective measured logatom intelligibility was 83.4%. In authors' opinion this was due to the lack of temporal synchronization between the two samples. It should be noted that the signals used for the test were not exactly the same what can caused the differences of particular parts (vowels and consonant) of speech signal structure.

In order to perform an objective evaluation of the quality of synthetic speech, the Opera (Voice/Audio Quality Analyzer) program was used, which is used for objective measurements of transmission quality in a telecommunications channel, in accordance with the ITU-T recommendation (P.862) for the PESQ method [22]. In this method, both the reference and test signals are compared in terms of delay and level, and then any discrepancies are compensated for, and the signals undergo a series of transformations. The final step is to compare the two signals using a cognitive model. The same synthesized speech signals as subjectively evaluated previously were used for the study, while the signals previously used to evaluate transmission quality in digital radio [26,27,29] served as reference samples. It should be noted that the synthesizers generated the same sentences as spoken by the voiceovers. As in the case of the logatom intelligibility study, no sufficiently reliable results were obtained in the trial phase. The results of the objectively measured quality evaluation did not exceed 2 MOS, while subjective tests for the best evaluated

synthesizers yielded values above 4 MOS. According to the authors' opinion, the reason for this, as in the case of measuring logatom intelligibility, is insufficient synchronization between the signals: tested and reference. For this reason, further attempts to use the tools used in transmission quality studies to assess the quality of synthetic speech were abandoned.

IV. DISCUSSION

On the base of the results of experiments performed some facts can indicate the difficulties of synthesized speech assessment in comparison to the speech quality evaluation for natural speech signals. In the case of synthetic speech, the generated logatoms sounded quite unambiguous, so listeners had only a little difficulty to identify them. This is reflected in the small standard deviation values of obtained results. The exception was one device (the WP synthesizer). The large value of the standard deviation of the results of the evaluation of the logatom intelligibility of the WP synthesizer may indicate the relatively high uncertainty of the evaluations given by the listeners due to the short time duration of generated vowels. It is also worth mentioning that during the training session, the listeners emphasized the lack of decisiveness in evaluating the logatoms generated by this particular device.

It should be noted that the obtained values of logatom intelligibility should ensure 100% sentence intelligibility for all synthesizers tested, since for Polish speech, logatom intelligibility as low as 25% guarantees 100% sentence intelligibility [30]. Therefore, the speech sound quality generated by the WP synthesizer was not rated the lowest. Other factors may also have influenced the quality rating, such as detecting the meaning of a sentence through context perception, or the impression of low artificiality of the generated sound, which, combined with the low distinguishability of logatoms, gave the impression of natural speech perception. It means that sentence intelligibility is a less crisp criterion than logatom intelligibility index, i.e. the brain "gets" for itself those things it does not hear. Moreover, sentence intelligibility does not need to be measured, because if we measure logatom intelligibility index, and the index is higher than 25% for Polish speech, it guarantees the high scores of the sentence intelligibility.

Discrepancies between speech intelligibility and MOS quality ratings also exist for Syntalk and eSpeak (both voices) and Dant Free synthesizers. It should be noted, however, that here the situation is the opposite of that for the WP synthesizer: low speech quality ratings here correspond to relatively high logatom intelligibility values. This may indicate a discrepancy of criteria in the evaluation of synthetic speech on the principle: something that is very clear is not evaluated naturally, and the naturalness of speech sound is one of the important components of subjective quality assessment [12,13,31].

Thus, the criteria for evaluating the quality of natural and synthetic speech can feature different weights, both for subjective and objective evaluation. So it became necessary to apply to objective measurements at the end of the measurement chain a block responsible for comparing signals according to the cognitive model. With subjective measurements, the element of cognitive comparison occurs, so to speak, on its own, through the perception of the speech signal and its interpretation.

In the case of objective assessment of synthetic speech, the difficulty in using transmission quality assessment tools seems to be not so much in the temporal synchronization between logatoms and words, but in the selection of the pattern. To assess transmission quality, a CD-like quality pattern is most often used and then its degradation is assessed. In contrast, when evaluating synthetic speech, the speech generated by the synthesizer is the pattern, so to speak. The need for suitable conditions has clearly been recognized in the field of prosody testing. For example, as for testing speech melody one can find that the baseline condition is synthesized on a monotone, constant pitch. In analogy with the random duration reference, a random melodic reference may be included for the sake of validation by making the pitch go up and down within reasonable limits.

Objective assessment of consists of getting a score to classify the measured system. The concept of objective evaluation of synthetic speech should be in accordance with the Quality of Experience criteria. Unfortunately, the objective metrics do not align well with human perception. For example, if the intelligibility of synthesized speech is satisfactory, it may still lack the naturalness of the human voice, its specific character and prosody. This is because this type of voice degradation in synthetic speech is different from that of a telecommunications chain, observed in low-bit-rate speech transmission and coding [18], or in electroacoustic channels at the high values of non-linear distortion [31]. This limits the use of objective measurement to system tuning, while the final evaluation must be based on a subjective listening test. Many of the accurate objective measurement methods require access to natural speech in order to comparison. Moreover, it seems to be a requirement to parametrize the natural speech signal according to the principles of fuzzy logic, which can partially account for aspects of prosody. The most common aspects to score in speech quality are intelligibility, segmental quality and prosodic correlates which includes the temporal changes of pitch of the voice as well as the voiced/unvoiced articulation and aspects of understanding and interpreting the spoken contents [32]. When trying to capture naturalness as the quality metrics, one should focus on spectral features and consider prosody as a secondary problem, an approach that seems to be based on a bias that is difficult to motivate from the phonetic point of view.

V. CONCLUSION

Modern technical and software solutions make it possible to achieve good synthetic speech parameters that affect logatom intelligibility as well as the clarity and recognition of speech. However, these parameters do not have to guarantee a good quality of synthetic speech. This is due to the lack of consideration in the process of its production of such features as naturalness and fluency of speech and temporal changes in intonation and articulation of sound. This suggests the continuation of work on speech synthesis in terms of considering prosodic features by AI-based algorithms, or deep learning non-intrusive speech assessment models with cross-domain features [33].

Considering the applicability of subjective and objective methods for assessing the quality of synthetic speech, it has been found that objective methods are not applicable at this stage. The fundamental reason for this is the lack of temporal synchronization between the test sample and the reference one. The second factor is the lack of a proper pattern, taking into account such aspects of human speech as the change in the melody of speech, its expression and articulation. Of course,

other parameters such as formant frequencies and the temporal structures of individual voices are also important in the process of speech recognition, and in assessing the quality of synthetic speech.

Despite the fact that subjective methods are more time- and cost-consuming, they will continue to be reliable tools used to assess the quality of synthetic speech for professional quality verifications.

REFERENCES

[1] H. Dudley, "The carrier nature of speech", *Bell System Technical Journal*, vol. 19, no 4, pp 495-515, 1940, https://doi.org/10.1121/1.1916020

[2] H. Dudley, "Fundamentals of speech synthesis". *Journal of the Audio Engineering Society*, vol. 3, no 4, pp 170-185, 1955.

[3] J. Benesty, M. M. Sondhi, Y Huang, (Eds.), "*Springer handbook of speech processing*," Berlin: Springer, 2008

[4] R. E. Donovan, "*Trainable speech synthesis".* Doctoral dissertation, University of Cambridge, 1996, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=63011431015dd41881987e37f138c4060975442c (07.08.2024)

[5] M. Stone, C. H. Shadle, "A history of speech production research." *Acoustics Today*, vol. 12, no 4, pp. 48-55, 2016, https://www.isca-archive.org/hscr_2019/hoffmann19_hscr.pdf (27.07.2024).

[6] T. Dutoit, T.. "High-quality text-to-speech synthesis: An overview". *Journal Of Electrical And Electronics Engineering Australia*, vol. 17, no 1, pp 25 - 36, 1997.

[7] S. R. Mache, M. R. Baheti, C. N. Mahender, C. N. "Review on text-to-speech synthesizer." *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no 8, pp. 54-59, 2015, https://doi.org/10.17148/IJARCCE.2015.4812

[8] S. Furui, "*Digital speech processing: synthesis, and recognition*." CRC Press. 2018, https://doi.org/10.1201/9781482270648

[9] F. Khanam, F. A. Munmun, N. A. Ritu, A. K. Saha, M. Firoz,. „Text to speech synthesis: a systematic review, deep learning based architecture and future research direction." *Journal of Advances in Information Technology*, vol. 13, no 5, pp. 398 – 412, 2022, https://www.jait.us/uploadfile/2022/0831/20220831054604906.pdf (07.08.2024), https://doi.org/10.12720/jait.13.5.398-412 ,

[10] X. Tan *et al*., "Natural Speech: End-to-End Text-to-Speech Synthesis With Human-Level Quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234-4245, 2024, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10409539, (07.08.2024), https://doi.org/10.1109/TPAMI.2024.3356232 .

[11] E. V. Raghavendra, P. Vijayaditya and K. Prahallad, "Speech synthesis using artificial neural networks," *2010 National Conference On Communications (NCC)*, Chennai, India, 2010, pp. 1-5, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5430190 (07.08.2024), https://doi.org/10.1109/NCC.2010.5430190 .

[12] V. J. van Heuven, R. van Bezooijen, "Quality Evaluation of Synthesized Speech", in: W.B. Kleijn, K.K. Paliwal, (eds.), *Speech coding and synthesis,* pp. 707-708, 1995. Elsevier, Amsterdam.

[13] N. Kitawaki, H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems", IEEE Communications Magazine, October, pp. 36 – 44, 1988.

[14] S. Brachmanski, "Selected problems of speech transmission quality assessment" (Wybrane zagadnienia oceny jakości transmisji sygnału mowy), Wrocław University of Science and Technology Edition, 2015, (in Polish).

[15] S. Brachmański, "Automation of subjective measurements of speech intelligibility in analogue telecommunication channels", *Archives of Acoustics*. vol. 33, no 3, pp. 341 – 350, 2008, https://acoustics.ippt.pan.pl/index.php/aa/article/viewFile/536/46726.07.2024).

[16] M. Daniluk, A. P. Pietrzak, "Comparative analysis of natural and synthesized Polish speech". *Int. Journal of Electronics and Telecommunication*, vol. *70, no* 2, pp. 361-366 2024, https://doi.org/10.24425/ijet.2024.149553

[17] Cooper, E., Huang, W. C., Tsao, Y., Wang, H. M., Toda, T., J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech." *Acoustical Science and Technology*, vol. 45, no. 4, pp. 12 - 24, 2024. https://doi.org/10.1250/ast.e24.12

[18] R. J. Beaton, J. G. Beerends, M. Keyhl, W. C. Treurniet, "Objective perceptual measurement of audio quality", In *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*. Audio Engineering Society, 1996.

[19] F. Holly, I. Scott, O. Eunmi, "Objective Measures of Voice Quality for Mobile Handsets", *140 AES Convention*. Audio Engineering Society, Paper 9532, 2016.

[20] M. Torcoli, T. Kastner, J. Herre, " Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, Vol. *29*, pp. 1530 - 1541, https://doi.org/10.1109/TASLP.2021.3069302

[21] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality.", Geneva, Switzerland, 1996

[22] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs", Geneva, Switzerland, 2001.

[23] ITU-T Recommendation P.863, "Perceptual objective listening quality assessment", Geneva, Switzerland, 2018.

[24] ITU-R Recommendation. 1534-1: Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," Geneva, Switzerland, 2001.

[25] S. Brachmanski, "Test material used to assess speech quality in Poland", in: *Acoustics, acoustoelectronics and electrical engineering*, F. Witos (ed.), Gliwice, pp. 65-79, 2021.

[26] S. Brachmanski, M. Kin, P. Zemankiewicz, "Subjective Assessment of the Speech Signal Quality Broadcasted by Local Digital Radio in Selected Locations in Wroclaw under Studio and Home Conditions", *Int. Journal of Electronics and Telecommunications,* vol. 68, no. 4, pp. 687 – 693, 2022, https://doi.org/10.24425/ijet.2022.141290

[27] S. Brachmanski, M. Kin, N. Rurzyńska, "Objective Assessment of the Speech Quality Broadcasted by Local Digital Radio in Selected Locations in Wroclaw", *Int. Journal of Electronics and Telecommunications,* vol. 70, no. 3, pp. 603 – 608, 2024, doi: 10.24425/ijet.2024.149585

[28] F. Holly, I. Scott, O. Eunmi, "Objective Measures of Voice Quality for Mobile Handsets", *140 AES Convention*. Audio Engineering Society, Paper 9532, 2016

[29] M. Kin, S. Brachmański, "Quality assessment of musical and speech signals broadcasted via Single Frequency Network DAB+". *Int. Journal of Electronics and Telecommunications*, vol. 66, no. 1, pp. 139 – 144, 2020, https://doi.org/10.24425/ijet.2020.131855

[30] W. Myślecki, W. Majewski, "Relations between subjective and objective measures of speech transmission quality evaluation", in: *Proceedings of 6th FASE Symposium,* Sopron, Budapest, pp. 137-141, 2-6 September 1986.

[31] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, "PEAQ-The ITU standard for objective measurement of perceived audio quality", *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3 - 29, 2000.

[32] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. LeMaguer, Z. Malisz, E. Szekely, Ch. Tannander, J. Voβe, "Speech synthesis evaluation – State-of-the-art assessment and suggestion for novel research program", in: *Proceedings of 10th ISCA Speech Synthesis Workshop*, Vienna, pp. 105 – 110, September 2019. https://doi.org/10.21437/SSW.2019-19

[33] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 54-70, 2022. https://doi.org/10.1109/TASLP.2022.3205757