Belarusian Language Oriented Intelligent Voice Assistants

Yauheniya Zianouka, Volga Dydo,
Maxim Lutich, Vitalij Chachlou, Juras Hetsevich
United Institute of Informatics Problems
National Academy of Sciences of Belarus Minsk, Belarus
{evgeniakacan, olgadydo1,
maksim.lyutich, vitalikhokhlov, yuras.hetsevich}@gmail.com

Vadim Zahariev, Veronika Krischenovich
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
{zahariev, krish}@bsuir.by

***Abstract*—The article represents Belarusian intelligent voice assistants platform in open access and free use. It depicts an architecture of question-answering systems, their versions and the principles of work. Also, Belarusian modern speech synthesis and recognition systems of new generation are described in detail, which are the core of AI-assistants. The platform employs a structured approach, including an input interface, data processing, and information search, to ensure relevant and accurate answers for users. The OSTIS technology is shown as a means of improving AI-assistants' responses.**

***Keywords*—artificial intelligence, question-answering system, natural language processing, voice assistant, text-to-speech system, speech recognition system, large language model, chatbot, OSTIS technology**

I. Introduction

Artificial intelligence (AI) is playing an increasingly significant role in the development of modern society. With the advent of new technologies and opportunities, the use of AI is becoming an integral part of our lives, influencing all areas of activity, from business and science to education and medicine. For example, question-answering systems use Natural Language Processing (NLP) techniques, speech synthesis and recognition systems, machine learning, dialogue systems and other algorithms to understand questions, search for relevant information and generate answers based on available data [1]. They solve the difficult task of understanding natural language, which is one of the key components of artificial intelligence.

Such question-answering systems as *Chat GPT-4, Midjorney, Google Gemini, GPT Yandex, Burd, Copilot, Mistral* are already well known. They process requests in different languages [2], with the exception of the Belarusian language. For Belarusian speakers, the Speech Synthesis and Recognition Laboratory of the UIIP of the National Academy of Sciences of Belarus has developed an interactive Voice AI-assistant platform [3] which contains a set of Belarusian-speaking female and male question-answering assistants. The concept of the platform is based on the provision of an effective and easy-to-use mechanism for performing general information and solving user problems in the Belarusian language [4]. Assistants are represented in three versions (Web-version, iOS and Android platforms for mobile applications, and chatbots on the Telegram social network). Each system is built using speech recognition and synthesis technologies, machine translation, and dialogue systems. They allow users to ask questions verbally or in text form and receive an audio/printed response quickly, with high quality and accuracy.

II. The architecture of intelligent voice assistants

The AI-assistants of the intelligent question-answering platform have the following structure (Fig. 1) [5]: 1. Input interface. The user can ask an assistant a question verbally or in text format. Voice requests are more difficult to process than text messages. Therefore, at this stage, *the Belarusian-language speech recognition system (BSRM)* [6] is used to convert the audio signal into text form for searching the most relevant response. The system processes voiced requests efficiently, which reduces the likelihood of an answer that does not match the request.

2. Processing of input data. The system performs a re-interpretation of the question to understand its meaning and context. This step includes a set of natural language processing tools that highlight the main semantic units in the question. If a chart with this user has been created before, a dialogue summary is loaded from the database for more accurate and related responses.

3. Information search, ranking of answers, and their extraction. The voice assistant is looking for suitable information that may contain the answer to the question in online resources. The GPT language model (namely the ChatGPT-3,5 model) is used in Belarusian-speaking question-answering systems. The model is pre-trained on huge sets of text data; therefore, GPT can generate text that makes sense, uses the correct grammar and sentence structure. If there are several possible answers, the system can use a ranking algorithm to select the most relevant one. ChatGPT-3.5 supports the Belarusian

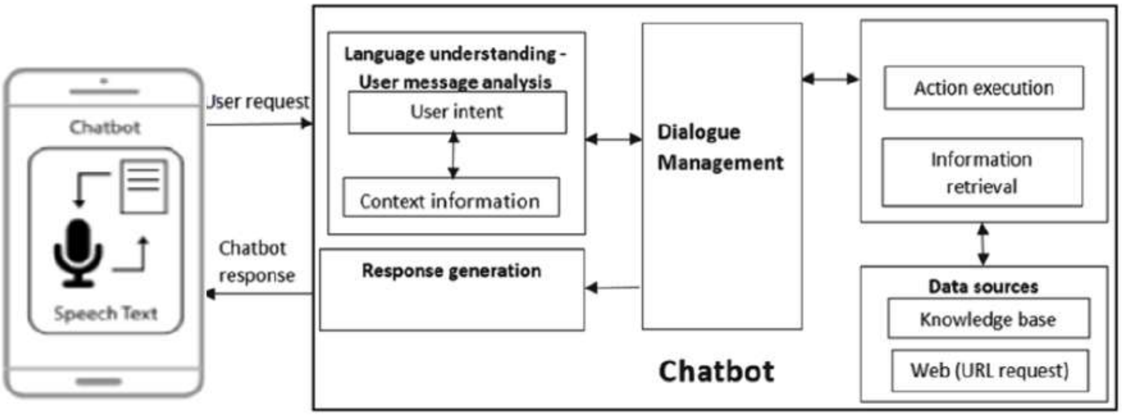


Figure 1. An architecture of intelligent voice assistant

language, but the quality of responses is not very good. To improve it, an additional machine translation unit has been developed, in which all queries are automatically translated from Belarusian to English using the Google Translate system. Then the most accurate answer given by ChatGPT goes back to the machine translation block, where it is already converted in Belarusian. Also, at this stage, the system saves answers to the language model in the database to create a summary of the dialogue with the user.

4. Forming a response. The voice assistant generates a response to the user in the form of a text or voice message, depending on the user's desire. An updated model of the Belarusian-language online speech synthesizer is used to output a voice message.

III. Belarusian speech recognition system using deep machine learning

To process the spoken question, a high-quality Belarusian speech recognition system (BSRS) [6] is used (fig. 2). It is based on an end-to-end architecture using deep learning and hosted on the Hugging Face platform, which allows users to create and share machine learning models and datasets.

To develop the BSRS, a large corpus of well-read texts in the Belarusian language was collected. The total duration of the audio recordings is 987 hours,voiced by 6160 speakers. The high variability of the collected data both from the point of view to speakers (gender, age, speech tempo, other features) and to recording conditions (various microphones, background noise, etc.) shows its quality. This is the first example of such big datasets for the Belarusian language.

To build a speech recognition model, a deep neural network architecture **wav2vec2** was chosen [7]. Its advantage is pretraining on a corpus of non-annotated data to study the ways of qualitative selecting features from the input recording. The obtained features are used for further subtasks, for example, to teach the model to convert speech into text. We used *"facebook/wav2vec2-base"* as the pre-studied model. Its further training was conducted on the Belarusian speech data set collected on *the Common Voice platform*. The training, validation and test samples were left unchanged; that is, the limit on the number of vocalizations of the same sentence was not removed, and the data set size was 345 909 audio recordings.

The speech recognition systems consist of 2 main components:

1) *The acoustic model* is a speech recognition system unit that builds a sequence of phonemes (or letters) that are pronounced with the greatest probability. It is based on the features selected from the input audio signal.
2) *A language model*, which is needed to translate a set of phonemes or letters obtained from an input audio recording into a set of the most likely words – the final transcription.

For the acoustic model, all audio recordings were converted to the following format: sampling rate 16 kHz, 1 channel (mono). We reduced text transcriptions to lowercase; removed all characters except letters of the alphabet and numbers; and replaced each sequence of characters-spaces (space, tab, etc.) by 1 space. We used *CTC* as a loss function to train the model for speech recognition tasks. The parameters were optimized using *the AdamW algorithm*, a corrected version of the popular Adam optimization algorithm. *the Gradient Checkpointing* method was also applied to optimize memory consumption. The selection of the best model was carried out using the*WER (Word Error Rate Metric)* on a validation sample [8].

The software implementation of acoustic model training was fulfilled on a popular framework for training NLP and ASR models *HuggingFace*. It is a shell over another framework — *PyTorch* [9]. The training was conducted on a server with 3 NVIDIA GeForce RTX

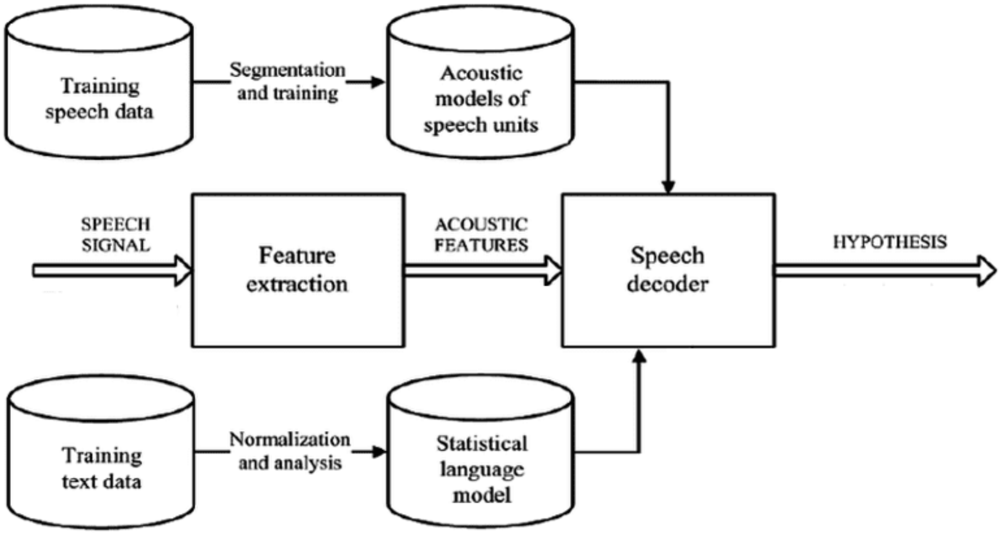


Figure 2. The structure of Belarusian Speech Recognition System

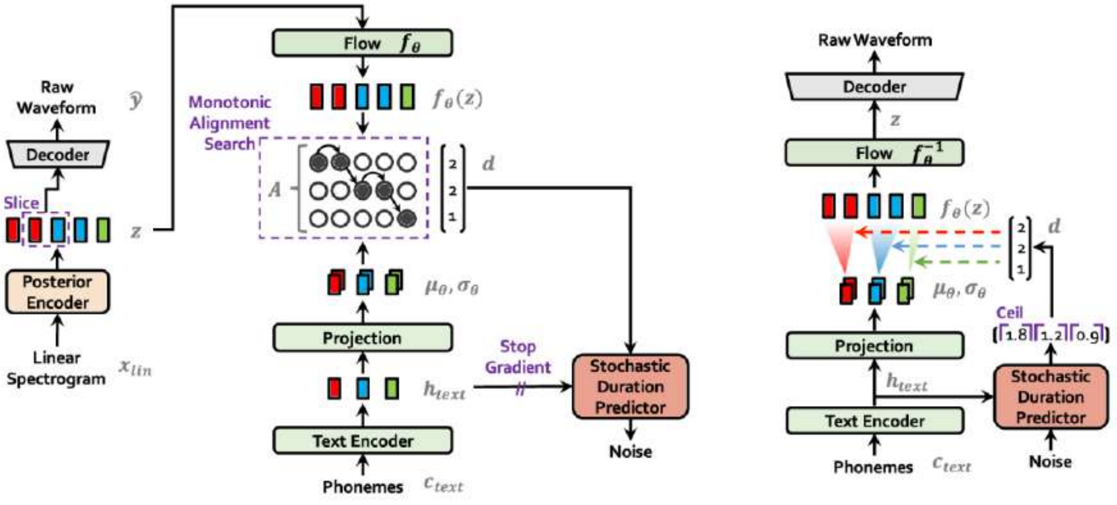
2080 Ti graphics cards. The size of the batch for training and evaluating the quality of the model was 48 (16 elements in the batch on each of the 3 video cards). The average time per epoch was about 8 hours. In this regard, and with acceptable metric values, training was discontinued after 5 epochs. Quality evaluation and compliance with intermediate parameters (checkpointing) were carried out several times during each epoch. To improve the predictions of the acoustic model, a 5-gram language model was constructed using modified *Kneser-Ney smoothing*. The popular *KenLM library*, created by the authors, was used for building such a model. The language model was trained on the text corpus described above. The total number of sentences needed to build a language model was 314 676. Adding a language model allowed for a reduction of WER from 0.187 to 0.124 in the test sample. The final result, WER 0.124 (or 12.4 per cent) is quite good for recognition models. For example, the current best value of test WER for the German Common Voice dataset is 5.7 per cent. Now the model is able to recognize arbitrary Belarusian speech at a fairly good level [10].

IV. A new generation Belarusian text-to-speech synthesizer

A new generation Belarusian-text-to-speech synthesizer is based on *The VITS (Variational Inference with adversarial learning for Text-to-Speech)* [11]. This is a one-stage non-autoregressive text-to-speech model capable of generating more natural sound than existing two-stage models such as *Tacotron 2, Transformer TTS, or even Glow-TTS*. Using a variational framework, VITS models a latent space of speech features, reflecting the inherent variability and uncertainty in speech generation. Competitive learning environment at VITS enhances the synthesis process. Collaborative learning involves training the discriminator network to distinguish between real and synthesized speech, while the generator network aims to produce speech that successfully fools the discriminator.

This adversarial interaction helps to improve the overall quality and naturalness of the synthesized speech samples. VITS serves as a stand-alone text-to-speech solution as it does not require a separate vocoder. The general architecture of VITS is depicted in fig. 3. It consists of *a Posterior encoder, a Prior encoder, a Decoder and a Stochastic Duration Predictor*. The Posterior Encoder and Decoder Discriminator modules are used only during training, not for speech output. For the Posterior Encoder, *16 residual WaveNet blocks* are used, consisting of layers of extended scrolls with an activation block and a communication pass. The back-end encoder takes *xlin* linearly scaled logarithmic spectrograms as input and produces latent variables Z with 192 channels. The idea behind the Posterior Encoder is to translate the audio data from the mel-spectrogram space to the normal distribution space. That is why the model uses a linear layer above the Posterior Encoder to obtain the average variance of the normal posterior distribution. Prior Encoder consists of *Text Encoder, Projection Layer, Normalizing Flow, and uses Monotonic Alignment Search (MAS)*. Like Posterior Encoder, Prior Encoder aims to map textual data from phoneme space to normally distributed space [12].

Collecting data for model training is an essential step for the development of a Belarusian text-to-speech system. Given the limited availability of specialized datasets for the Belarusian language, the *CommonVoice dataset* was chosen as the main data source. It is a large collection of voice recordings collected from voluntary participants who read sentences in different languages.



(a) Training procedure

(b) Inference procedure

Figure 3. The structure of VITS

This dataset is available for free use and distributed under an open license, which makes it a valuable source for the development of linguacoustic resources in the Belarusian language. The disadvantage of the dataset from CommonVoice is that it is intended for speech recognition tasks and not specifically for synthesis. In this regard, the recordings from the dataset can be unprofessional and contain various noises or flaws that negatively affect the quality of speech synthesis.

To fix this, pre-filtering and selection of recordings were performed to ensure that the selected data were suitable for the speech synthesis task and of good quality. An audio material (about 20 000 words) met the selection criteria of the largest amount of available audio and a relatively low level of noise and artefacts. The analysis of the data was carried out with the aim of obtaining statistical information and understanding the peculiarities of the Belarusian language in the context of speech synthesis. It included an assessment of the distribution of phonetic units, the length of phrases, and other characteristics that may affect the quality and naturalness of synthesized speech.

The VITS model was trained using the *Coqui TTS library*, a popular open-source toolkit for TTS. Coqui TTS provides a complete set of tools and utilities for training and deploying TTS models [13]. The VITS model used the *Weights and Biases recorder* during the training process. It is a platform for tracking and visualizing machine learning experiments. It allows researchers and developers to log in and track training progress, metrics, and model performance in real-time.

By leveraging the capabilities of the Coqui TTS library and the integration of the Weights and Biases register, the VITS model was trained using robust TTS development toolkit. The use of Coqui TTS and the Wandb logger facilitated efficient experimentation, model optimization, and performance monitoring throughout the learning process. A server with an *Nvidia RTX4090 graphics card* was used for this. Parameters were optimized using the textitAdamW algorithm, a corrected version of the popular Adam optimization algorithm. The batch size for model training and evaluation was 74, and the model training time was 72 hours.

With the help of neural networks, the training and learning of the acoustic database were carried out. Therefore, it is created automatically and has quite accurate results. However, this synthesizer is large in size, which can reduce the speed of text output. The lack of processing numbers, figures, dates, and abbreviations are also considered a big drawback, which is the object of development for new methods and algorithms to correct this bug.

V. Versions of intelligent voice assistants

Voice assistants are available on the official website of the platform, the interface of which is presented in Belarusian, English, Russian and Chinese [14]. The user can choose a personal virtual interlocutor *(AIAlesBot, AIAlesiaBot, AIAlenaBot, AIBorisBot, AIKirylBot, AsistentBot)* and chat on the Internet by selecting the Web version or a smartphone by installing the application on Android or iOS operating systems. Question-answering systems are also available in the form of chatbots in Telegram Messenger. Launching an official website, the user chooses a convenient version, and then he is redirected to the version he has chosen, and enters a request. In addition, everyone can chat with chatbots (AIVasil, AIVasilina), which are narrowly focused on various fields of activity (general assistant, architect, business analyst, financial consultant, recruiter, project manager, legal adviser, marketer, engineer, programmer, teacher, and writer). The main feature of all voice assistants is that they process text or audio responses only in Belarusian, whereas text requests can be sent in English. Voice queries are also recognized only in Belarusian.

Currently, mobile applications for iOS and Android versions are being developed. The process of their creation consists of several stages: approving the design concept of the application and its functional features; work on the back-end; direct implementation of the project. The iOS mobile application is realized in *Swift* using the *UIKit framework*. This is a high-quality combination of technical solutions and a design concept. Technical tools include *the Massage Kit library* for creating a functional chat, as well as the integration of methods with the native *AVFoundation library* for flawless work with audio files. *The UICollectionView class* was selected to build the interface.

The writing language of the Android version is *Kotlin*. The following technologies were also used: Android Architecture, MVVM Architecture (using ViewModels), WorkManager, Kotlin coroutines, Java Threads, OkHttp, SQLite database (Room technology), Canvas.

The applications work as follows. The logo or the start video is loaded on the first page, and then the user can go to the settings or contacts screen. To select a certain assistant, the user must click on the assistant card, and then the chat window opens. Then a chatbot greets the user, after which the second one enters a text or voice request. After entering a request, he cannot set another request until he receives a response from the bot. The system automatically voices the message in the manner of the selected speaker.

The verification of question-answering systems takes place daily. According to statistics, chatbots are good at answering simple questions like *"How many colours are in a rainbow?", and "How much is 1+1?")*, as well as difficult ones (for example, *"Tell me about the very first film", "I need an interesting story about Shakespeare"*). Text queries and responses in English and Belarusian are quite high-quality. If the question was asked in Russian or some other languages, the bots will answer in English. On average, the user can receive a response from the chatbot within 10-30 seconds, which is a good result of the server side.

VI. Application prospects of OSTIS technology for voice assistants

Semantic technologies, particularly the OSTIS technology, offer numerous advantages when applied in the development of voice assistants tailored for specific languages, like Belarusian. Firstly, they provide a flexible framework for constructing dialogue rules, enabling efficient handling of user queries in natural language. This flexibility ensures adaptability to diverse conversational contexts, resulting in more intuitive and user-friendly interactions [16].

Secondly, semantic technologies, including OSTIS, offer mechanisms to limit the size and context of responses, addressing the issue of information overload and reducing irrelevant or excessive outputs. This targeted response approach enhances the user experience by delivering concise and relevant information.

Moreover, the application of such technologies contributes to reducing hallucinations, a common challenge faced in language generation models like GPT and other LLM's. By leveraging semantic understanding and contextual awareness, OSTIS-based voice assistants can generate responses that are more coherent and accurate, minimizing nonsensical or misleading outputs.

The integration of OSTIS technology also facilitates the development of intelligent voice assistants capable of capturing and analyzing user profiles and dialogue histories. This feature enables personalized interactions, where the assistant can tailor responses based on individual preferences and past interactions, thereby enhancing user satisfaction and engagement.

Furthermore, semantic technologies allow for seamless integration of domain-specific knowledge with general knowledge bases within voice assistant systems. This integration involves efficient handling of domain-specific queries and contextually relevant responses, ensuring the assistant's effectiveness across diverse topics and applications.

VII. Conclusion

The article depicts Belarusian question-answering systems, which are available on the AI-assistant platform. The goal of these assistants is to provide easy access to information in the Belarusian language. With the use of artificial intelligence, question-answering systems provide quick and accurate responses on various topics, including scientific discussions and entertainment suggestions. To do this, assistants invoke such technologies as text-to-speech and recognition systems, effective algorithms of search, and machine translation.

The relevance of the assistants is due to the lack of competitive chat-bots that support the Belarusian language, whereas there are a huge number of question-answering systems for other languages. Such developments make them more accessible to local users, offering a convenient option for searching for information and communication in a modern global Internet network and using computer technologies in their native language. Future plans for the platform include new functions such as creating custom bots, gathering user feedback, saving message history, supporting developers, and extracting information from the internet to enhance effectiveness.

Applying semantic technologies, specifically OSTIS, in the development of voice assistants brings numerous benefits, including enhanced dialogue management, reduced hallucinations, and personalized interactions. This integration not only improves response relevance and coherence, but also underscores the significance of semantic technologies for user satisfaction and engagement.

References

[1] Bansal,H.,Khan,R. A review paper on human computer interaction.Int.J.Adv.Res.Comput. Sci. Softw. Eng. 8, 53 (2018). Avaliable at: https://doi.org/10.23956/ijarcsse.v8i4.630 (accessed 11.05.2023).

[2] Adamopoulou, Eleni Moussiades, Lefteris An Overview of Chatbot Technology. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Avaliable at: https://www.researchgate.net/publication/341730184 (accessed 17.02.2024).

[3] AI Voice Assistant. Avaliable at: https://asistent.io/ (accessed 09.01.2024).

[4] Hiecevic,Ju, Zianouka, Ja., Dydo, V. Liucic, M., Pavucina, M. Bielaruskamounaja halasavaja pytalna-adkaznaja sistema [The Belarusian-language voice question-answering system]. Science and innovation. Minsk. 2023. №7 (245). P. 13-16. (In Belarusian)

[5] Hiecevič, J.S., Zianoǔka J.S., Trafimaǔ A.S., Bakunovič A.A., Latyševič D.I., Drahun A.JA., Sliesarava M.M., Tukaj M.S., Komplieks srodkaǔ realizacyi zadač štučnaha inteliektu dlia bielaruskaj movy [A complex of means to implement tasks of artificial intelligence for the Belarusian language]. Piervaja vystavka-forum IT-akadiemhrada Iskusstviennyj intielliekt v Bielarusi. UIIP NASB. Minsk. 2022. P. 64-73. (In Belarusian)

[6] Belarusian Speech Recognition. Avaliable at: https://corpus.by/BelarusianSpeechRecognition/?lang=en (accessed 31.09.2023).

[7] Dhankar, Abhishek Study of deep learning and CMU sphinx in automatic speech recognition. 2017 International Conference on Advances in Computing, Communications and Informatics,2017, pp. 2296-2301.

[8] Graves Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369-376.

[9] Heafield Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn Scalable modified Kneser-Ney language model estimation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2013, pp. 690-696.

[10] Trafimau, A. S. Autamatycnaje pierautvarennie bielaruskaha mauliennia u tekst [Automatic conversion of Belarusian speech into text]. XXI International Scientific and Technical Conference "Development of Informatization and State System of Scientific and Technical Information RINTI-2022", UIIP NASB, Minsk. 2022.P. 241-245. (In Belarusian)

[11] Jaehyeon Kim, Jungil Kong, Juhee Son Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Avaliable at: https://arxiv.org/abs/2106.06103 (accessed 03.02.2023).

[12] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Avaliable at: https://arxiv.org/abs/2010.05646 (accessed 13.02.2023).

[13] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu FastSpeech: Fast, Robust and Controllable Text to Speech. Avaliable at: https://arxiv.org/abs/1905.09263 (accessed 17.03.2023).

[14] Dydo, V. V., Liucic, M. S., Pavucina, M. A., Drahun, A. Ja., Chachlou, V. A., Trafimau, A. S., Zianouka, Ja. S., Hiecevic Ju. S. BIELARUSKAMOUNY HALASAVY AI-ASISTENT [BELARUSIAN-SPEAKING VOICE AI ASSISTANT]. XXII International Scientific and Technical Conference "Development of Informatization and State System of Scientific and Technical Information RINTI-2022", UIIP NASB, Minsk. 2023. P. 190-194. (In Belarusian)

[15] Golenkov V., Gulyakina N. A., Shunkevich D. V. Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems. Bestprint [Bestprint]. Minsk. 2021.

[16] Zahariev, V., Shunkevich, D., Nikiforov, S., Azarov, E. Intelligent Voice Assistant Based on Open Semantic Technology. Open Semantic Technologies for Intelligent System. OSTIS 2020. Communications in Computer and Information Science, vol 1282. Springer, Cham. 2020, pp. 121-145.

ИНТЕЛЛЕКТУАЛЬНЫЕ ГОЛОСОВЫЕ АССИСТЕНТЫ, ОРИЕНТИРОВАННЫЕ НА БЕЛОРУССКИЙ ЯЗЫК

Зеновко Е., Дыдо О., Шуст М., Хохлов В., Гецевич Ю., Захарьев В., Крищенович В.

В статье представлены интеллектуальные голосовые ассистенты на белорусском языке, размещенные на Интернет-платформе Voice AI-assistant. Главной целью разработки ассистентов является обеспечение эффективного и простого в использовании механизма предоставления общей информации и решения вопросов пользователей на белорусском языке. Вопросно-ответная платформа "Голосовой ИИ-ассистент"позволяет пользователю задать вопрос на белорусском языке текстовым или голосовым сообщением и получить на него звуковой или напечатанный ответ. За счет использования искусственного интеллекта она дает возможность получать быстрые, качественные и точные ответы по различным темам.

Ассистенты представлены в трех версиях (Web-версия, iOS- и Android-платформы для мобильных приложений и чат-боты в социальной сети Telegram). Каждая система построена с использованием технологий распознавания и синтеза речи, машинного перевода и диалоговых систем. Описанные в работе белорусскоязычные системы синтеза и распознавания речи свидетельствуют о высоком уровне развития речевых технологий на белорусском языке.

Актуальность данных ассистентов обусловлена отсутствием конкурентоспособных чат-ботов, поддерживающих белорусский язык, тогда как для других языков существует огромное количество голосовых ассистентов. Разработка устройств на белорусском языке делает их более доступными для белорусскоязычных пользователей, предлагая удобный вариант поиска информации и коммуникации в современной глобальной интернет-сети и использования компьютерных технологий на их родном языке.

Received 14.03.2024