Uzbek Speech Synthesis Using Deep Learning Algorithms

M. I. Abdullaeva[1(✉)], D. B. Juraev[2(✉)], M. M. Ochilov[2(✉)], and M. F. Rakhimov[2(✉)]

[1] Department of Computer Systems, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
malika.ilkhamovna@gmail.com

[2] Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad al-Khorazmi, Tashkent, Uzbekistan
dilsamtuit@gmail.com, raximov022@gmail.com

Abstract. This paper presents modern architectures for effective speech synthesis. Since each language has its own subtleties, the task of applying the world methods for the Uzbek language was relevant, due to the lack of research in this direction. The paper presents a method consisting of the acoustic model Tacotron and the neural vocoder parallel waveGAN. The formed speech corpus with the volume of 31 h of Uzbek speech is described. The quality of the synthesized speech was evaluated using the MOS scale, according to that the intelligibility and accuracy of the synthesized speech was 4.36 points out of five.

Keywords: neural vocoder · tts system · Tacotron · parallel waveGAN · speech corpus

1 Introduction

Text-to-speech (TTS) is the computer simulation of human speech from a textual representation using machine learning techniques. The first speech synthesis system, called "têtes parlantes" (talking heads), appeared in the 18th century and was a pioneer, but was an imperfect imitation of the human voice.

There are many speech synthesizers around the world that synthesize electronic texts written in different languages into speech signals, and all the methods and tools used in their synthesis differ in the lexical and phonetic features of the chosen language. Currently, one of the important issues is to conduct research on the organization of speech synthesis and preprocessing of electronic texts, for full and perfect linguistic expression of the existing features of the chosen language, as well as to achieve synthesis of speech signals close to the natural pronunciation.

Automatic speech synthesis technology may be useful in a variety of industries and areas, such as telecommunications, mobile devices, industrial and consumer electronics, automotive industry, educational systems, computerized

H. Zaynidinov et al. (Eds.): IHCI 2022, LNCS 13741, pp. 39–50, 2023.
https://doi.org/10.1007/978-3-031-27199-1_5

systems, Internet services, access restriction systems, aerospace industry, the military-industrial complex. Speech synthesis technology offers great opportunities for people with physical disabilities. Speaking machines have been developed for the blind and visually impaired. For the dumb there are portable speech synthesis devices in which a message is typed on a keyboard, allowing communication with other people.

Nowadays, research on the recognition and processing of Uzbek speech is being carried out, and many scientific papers have been published on the results of this research. However, studies on the synthesis of Uzbek electronic texts into speech signals and the development of computational linguistics for the Uzbek language are insufficient. In particular, there is a lack of research on text analysis and processing, syllabic representation of texts, detection and correction of grammatical errors in the text, and real-time speech synthesis systems.

This research paper describes the general scheme of TTS systems, its constituent stages, their description and sequence. The classification of methods of text-to-speech conversion systems is given. In this paper defined families of acoustic models and given the capabilities and goals of neural vocoders. The paper includes the proposed method as a sequence of application of Tacotron 2 as an acoustic model and Parallel WaveGAN as neural vocoder. The Tacotron 2 and Parallel WaveGAN architectures are also described below. The peculiarity of the proposed method is that this sequence is applicable to the synthesis of Uzbek speech with its peculiarities.

2 Related Works

TTS system overcame great development and the first synthesizers were mechanical, which could generate separate sounds or small fragments of fused human-like speech like musical instruments [2]. Due to the age of the field there is a large number of methods of speech synthesis (Fig. 1). In scientific papers [1–4] are detailed reviews on existing synthesis technologies, where described their advantages and disadvantages, as well as their clear differences from each other.

The most simple and yet effective method is concatenative method. In works [1,5,6] the concatenative method algorithm is described and given its results in solving speech synthesis problems. This method is based on combining segments of recorded speech. The most important disadvantage of the method is the need for a large storage and the inability to apply various changes to the voice.

Studying the works [1,3,5,6,16] it can be concluded that speech synthesis can be achieved on the basis of a small amount of data. On the other hand, speech synthesis based on the selection of units proved that it is possible to reconstruct all the nuances and characteristics of the voice, if there is a large database. Hence, combining the two HMM synthesis and unit selection-based synthesis methods in one hybrid approach is another solution and method for high-fidelity speech synthesis [23,24].

The most advanced methods are those that are based on deep learning [11,15, 17,18]. Such speech synthesizers, are trained on recorded speech data. Common

deep learning based synthesizers are WaveNet from Google DeepMind, Tacotron from Google, and DeepVoice from Baidu.

One of the modern and high-quality acoustic models is the Tacotron model from Google. In [15,17] the method based on the acoustic model of the Tacotron2 family is described. The key points of the method, which are worth paying attention to, are given. General architecture of the acoustic model Tacotron2 is described too.

In scientific articles [12,18], there is information about the Parallel WaveGAN neural vocoder, which is the final step in modern TTS systems. This vocoder converts the acoustic features, which is received at the input of the acoustic model, into a speech signal.

3 Description of Modern TTS Systems

Speech synthesis is a long-developing area and in the process of development it has opened many methods that differ from each other as the quality of the synthesized speech, as well as the complexity of the algorithm, the amount of memory occupied. Figure 1 shows and classifies the basic and most effective methods of speech synthesis.

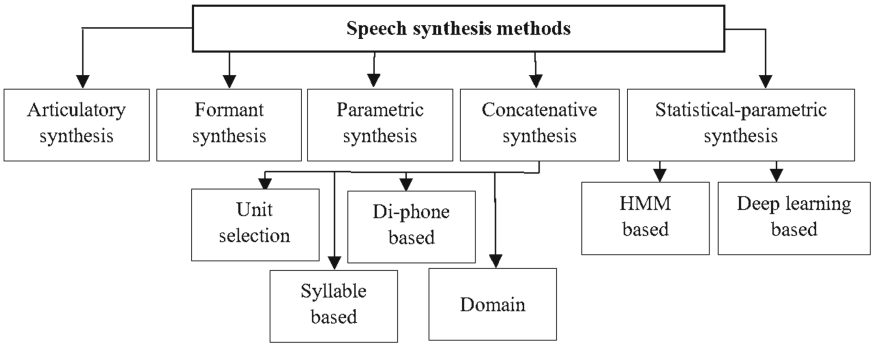

Fig. 1. Classification of text-to-speech synthesis methods

Speech synthesis technology in modern TTS systems consists of stages front-end and back-end, which in turn consist of a number of steps (Fig. 2). Each of these steps are described below [7].

Linguistic Analysis. This stage consists of text preprocessing and the main task is normalization of non-standard words. Normalization is the process of identifying numbers, abbreviations, acronyms, and idioms and converting them to full text, usually based on the context of the sentence.

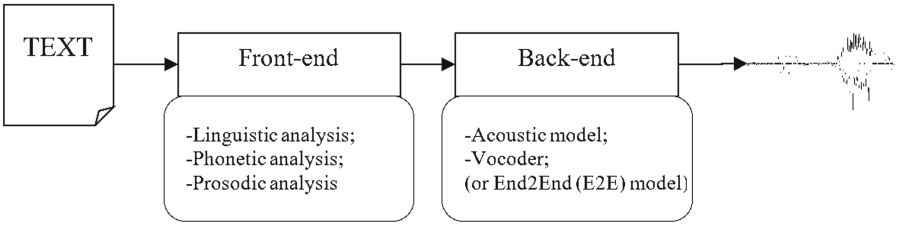

Fig. 2. General scheme of TTS systems

Phonetic Analysis. This step is a conversion of a grapheme into a phoneme. It is known that the grapheme is the minimal unit of writing and the phoneme is the minimal unit of oral speech.

Prosodic Analysis. At this stage, the boundaries of syntagms, localization and duration of pauses are determined, and the intonational type of phrases and the place of phrase emphasis in them are selected.

The existing methods for choosing the place of pauses can be divided into the following groups

1. Determining the places of pauses and boundaries of syntagmas according to the rules.
2. Determination of the place of pauses with the help of full parsing of sentences.
3. Determining the location of pauses using statistical methods.

Determining the Intonational Type of Syntagmas and the Place of Phrasal Stress. At this step, intonation transcription can be performed: the intonational type of syntagmas and the place of phrasal and emphatic stress are determined. Depending on the system of intonational transcription adopted, the rules may be more or less complex, but, in general, they are based on the analysis of punctuation marks (the simplest option) or the use of full/partial syntactic and semantic analysis of the sentence. Various statistical methods can also be used, for the training of which a text base is required, pre-marked with intonational transcription.

In modern TTS systems, the back-end environment is a synthesizer. It generates speech by converting each unit of transcription into sound using a selected method, algorithm or vocoder [15–17]. Thus Back-end consists of an acoustic model and a neural vocoder to approximate the parameters and relations between the input text and the waveform that constitute speech (Fig. 3).

Acoustic Model. The acoustic model algorithms are optimized to convert the pre-processed/normalized text into Mel spectrograms, thus converting the vector of linguistic features into acoustic features [19,20]. It is known that the spectrogram ensures that all significant sound features are taken into account and

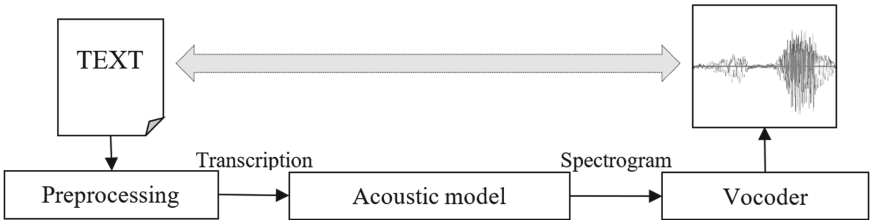

Fig. 3. Block diagram of the two-stage TTS system

carry high-level features. It is on the mel spectrogram that accents, features of interphoneme transitions and speaker's pronunciation are determined. In TTS systems the acoustics, with which the speech case is assembled, plays a very important role. For today there are various speech corpus in open access for the world languages. The most famous and quality ones are listed below:

- LJ Speech - EN, single speaker, 24 h
- Libri-TTS - EN, multi-speaker, 585 h
- RUSLAN - RU, single speaker, 29 h
- NATASHA - RU, single speaker, 13 h
- M-AILABS - multi language, 1000 h, 47 h of Russian

There are two main types of acoustic models - Tacotron family and Fast family.

Neural Vocoder. The input data for the latter stage are Mel spectrograms, which are converted into a waveform using a neural vocoder. Although there are many different types of neural vocoders, among them a special place belongs to vocoders with GAN(Generative Adversarial Networks) basis. For example, Parallel WaveGAN, Multi-band MelGAN, HiFiGAN, Style MelGAN.

Vocoders with GAN basis are based on a generator and a dicriminator, between which there is a constant interaction and struggle. The purpose of the generator is to generate high-quality speech, which will be close to the natural one, and the discriminator is focused on whether the generated speech is natural from the speech corpus or generated by the generator.

4 Method Description

The technology applicable for speech synthesis of the world languages, alas, is not suitable for the Uzbek language. This is due to the peculiarities of the language, such as unique letters, syllables and words. For this reason, the task of speech synthesis for the Uzbek language is relevant and unsolved to this day.

To develop the method of Uzbek speech synthesis, we have proposed using Tacotron2 architecture as an acoustic model for transition from transcription to

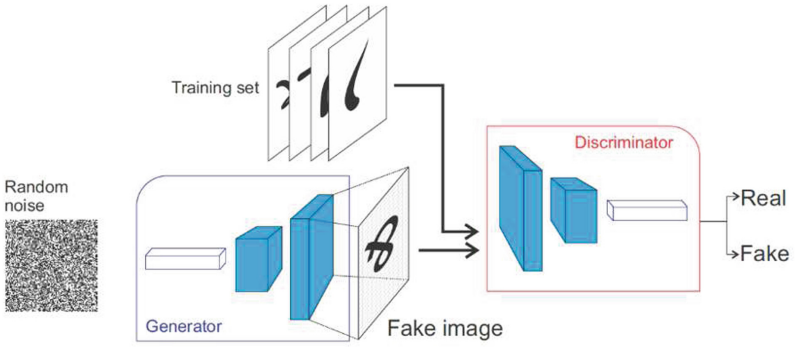

Fig. 4. Neural vocoder architecture based on Generative Adversarial Networks

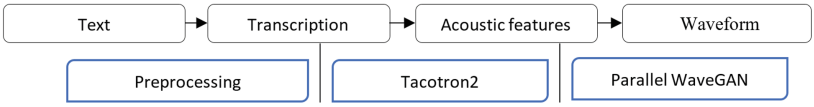

Fig. 5. General scheme of the sequence of operations for the Uzbek speech synthesis system

mel spectrogram, and Parallel WaveGAN architecture as a neural vocoder for mel spectrogram vocalization, as shown in Fig. 5.

Consider the architecture of Tacotron2. In architecture Tacotron2 has encoder, which is designed to work with embedding phonemes, has decoder with two heads - predicts next mel and Stop Token. Stop Token learns a binary classification of whether to stop producing speech. The output of the other linear projection goes to 3 points: in post-Net, residual connection and pre-net, which consists of two layers with dropout.

Parallel WaveGAN has a WaveNet generator and differs from the original WaveNet in that:

- Uses non-causal convolutions instead of causal convolutions;
- Takes random noise as input.
- The model is not autoregressive.

The detailed architecture of a simple and efficient method of generating parallel signals based on the Parallel WaveGAN is shown in Fig. 7.

Assessing the difference between the features of true and generated speech presents a particular difficulty. Without introducing such an estimate into the loss function, the convergence process will be extremely slow and unstable [22]. To solve this problem, Multi-resolution STFT loss functions were first proposed, which for brevity we will call STFT loss.

Let x is audio and $\hat{x}$ the generated one corresponding to x mel-spectrogram.

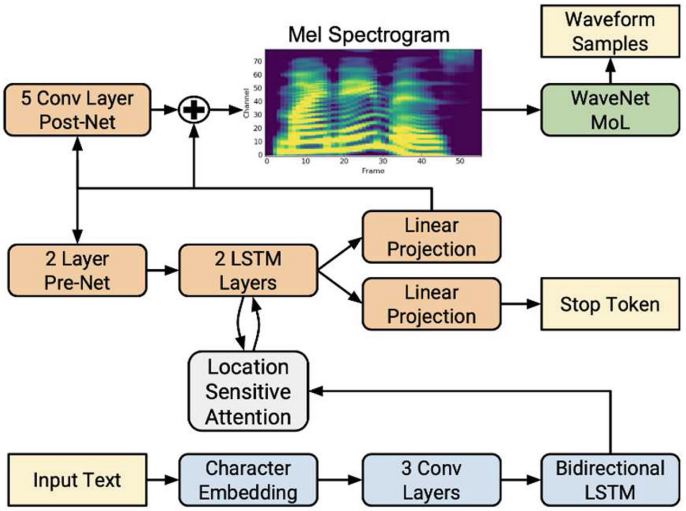

Fig. 6. Tacotron2 architecture

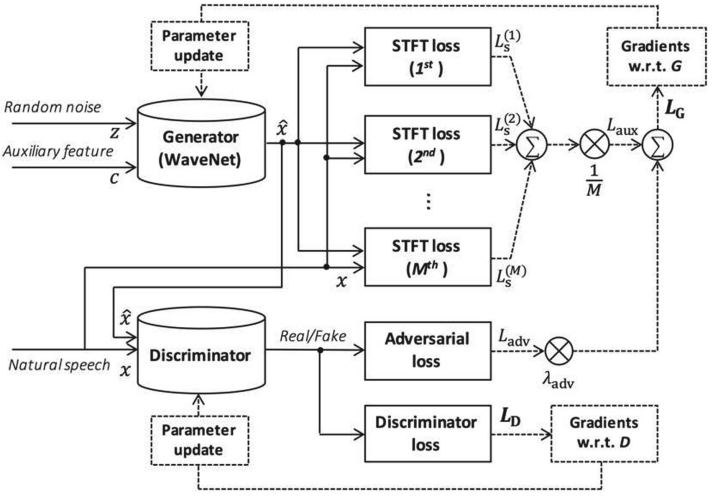

Fig. 7. Parallel WaveGAN architecture with STFT losses

Then for some chosen uniform STFT parameters:

$$L_{sc}(x, \hat{x}) = \frac{|||STFT(x)| - |STFT(\hat{x})|||_F}{|||STFT(\hat{x})|||_F}$$

$$L_{mag}(x, \hat{x}) = \frac{1}{N}||\log|STFT(x)| - \log|STFT(\hat{x})|||_1$$

where $\|.\|_F$- denote the Frobenius and $\|.\|_1 - L_1$ norms;

To increase the variety of patterns and scales of structures that are involved in a given loss function, we sum these loss functions as for different sets of STFT parameters:

$$L_{STFT} = \underset{x,\widetilde{x}}{E}\left(\frac{1}{3M}\sum_{m=1}^{M}\sum_{p=1}^{3} L_{sc}^{m}\left(x_p,\widetilde{x}_p\right) + L_{mag}^{m}\left(x_p,\widetilde{x}_p\right)\right)$$

5 Speech Corpus

The main component of modern high-quality TTS systems is a speech corpus (voice corpus) with a large volume [8]. The speech corpus (SC) is a set of a large number of audio data and their textual transcriptions. Tacotron2 learns language features from the speech corpus. In practice, the quality of synthetic speech depends on the quality of the speech corpus [9,10,21].

The most common methods of RC formation for TTS systems are:

- Recording of the speaker reading a pre-prepared text material;
- Recording of the speaker saying spontaneous speech, narratives, etc.

Both methods are expensive because of the need to involve additional specialists and speakers for pre-processing of text information and post-processing of transcriptions and corresponding audio data. Nevertheless, the first method has the advantage of being able to adapt the TTS system being developed to a particular domain by incorporating terminology and sentences from that domain into the SC.

6 Experiments and Results

The Aim of the Research Work. To apply the above method to synthesize Uzbek speech with high accuracy.

Description of the Speech Corpus

Within the scientific work the speech corpus was formed. The total volume of the Uzbek speech corpus for the speech synthesis systems was 31 h. This volume of the speech corpus was voiced by two speakers separately. The speech signals were recorded in a studio environment in .wav audio format with a sampling rate 22050 Hz, quantization of 16 bits and mono type.

A total of over 14,523 utterances were used in the texts provided for reading. There are 170 thousand words in the sentences, and 25 thousand of them are not repeated words in Uzbek.

Using the above proposed algorithms of linguistic analysis, the experts checked the quality of each statement and compliance of the statements with the audio data. This expert procedure was conducted manually. As a result, only verified transcriptions and their quality audio soundings were stored in the speech corpus.

Statistics of the speech corpus based on the parameter length of utterances in seconds vs the number of utterances with the current length in the speech corpus is shown in Fig. 8. According to Fig. 8, we can argue that the speech corpus mostly consists of utterances of 10 s and similar utterances in the speech corpus 794 audio. And the most rare were expressions with 39 s, which in the speech corpus in total 13.

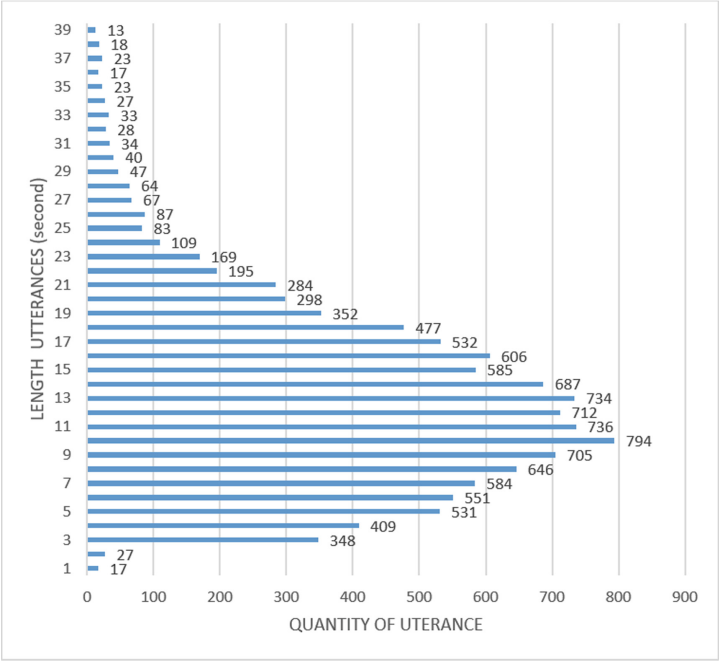

Fig. 8. Statistics of the Uzbek speech corpus

Description and Training of the Acoustic Model Tacotron2

The scientific work uses logarithmic spectrograms with a Hann window, a frame length of 50 ms, a frame shift of 12.5 ms, and a Fourier transform of 2048 points. In the paper, the sampling rate is defined as 22 kHz.

The Tacotron 2 was trained using the word sequence as input and the mel spectrogram extracted from the recorded speech. The model contained 5 encoder layers and 8 decoder layers.

The model was trained on an NVIDIA DGX-2 server with a 32G NVIDIA TESLA V100 GPU. Figure 9 shows the results of training on the Tacotron2 model. The trained data were pairs of audio data and their transcriptions from the speech case. For the speech corpus with a volume of 31 h.

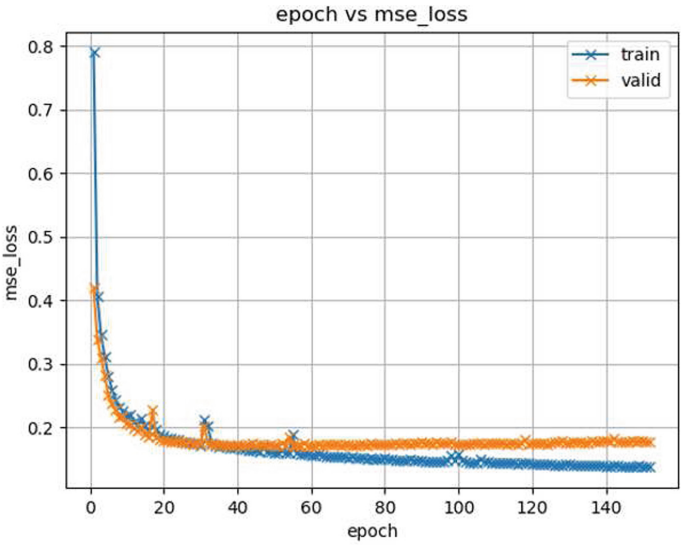

Fig. 9. Training results of the Tacotron2 model based on the developed speech corpus

6.1 Evaluation of the Developed TTS System

There are various methods of evaluation of synthesized speech, unfortunately, all of them are of subjective type. The following are the evaluation methods for TTS systems:

- Mean opinion score (MOS).
- MUSHRA
- Side by side SBS
- Robotness

Among them, MOS is particularly widely used, because of the availability of the evaluation method in different understandings.

To evaluate developed TTS system 12 Uzbek linguists (further experts) from Uzbek language department were involved. Each expert was given synthesized signals by the developed system, as well as a textual representation of the audio data. The synthesized speech was rated from 1 to 5, with 1 being very poor, 2 being unsatisfactory, 3 being satisfactory, 4 being good, and 5 being excellent. Finally, all the expert opinions were compiled into a table and an analysis was performed on them.

Table 1. MOS results for the proposed method

Model	MOS
Tacotron2 + Parallel WaveGAN	4.36 ± 0.09
Recording	4.72 ± 0.01

When testing the applicability and quality of the proposed combination of Tacotron 2 as a model to calculate the acoustic parameters and Parallel WaveGAN for speech synthesis, the results were obtained 4.36 ± 0.09. This figure is high enough, and the proposed method is applicable to the synthesis of Uzbek speech.

7 Conclusion

The results of the synthesized speech showed that the synthesis of Uzbek speech is achieved with a score of 4.36 (excellent) according to the MOS evaluation method. These scores were achieved using the Tacotron2 acoustic model and the Parallel WaveGAN neural vocoder.

While training the acoustic model of the speech corpus, it became known that the quality of the synthesized speech is fully dependent on the quality of the voice corpus. In order to form a high-quality Uzbek speech corpus, a method was developed that includes text preparation, audio recording, text-to-audio synchronization, and a final check to match the audio data to their transcriptions. According to the results of MOS we can state that the chosen method of speech corpus formation is effective and efficient. In addition to assessing the quality of the synthesized speech, the feature and importance of STFT loss and its impact on the quality of the synthesized speech were identified.

References

1. Kireev, N.S., Ilyushin, E.A.: Review of existing algorithms for text-to-speech conversion. Int. J. Open Inf. Technol. **8**(7) (2020). ISSN: 2307-8162
2. Rybin, S.V.: Textbook on the Discipline "Speech Synthesis". SPb: ITMO University, 92 p., (2014)
3. Prahallad, K.: Automatic building of synthetic voices from audio books. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (2010). CMU-LTI-10-XXX, July 26
4. Allen, J., Hunnicutt, M.S., Klatt, D.H., Armstrong, R.C., Pisoni, D.B.: From Text to Speech: The MITalk System. Cambridge University Press, NY (1987)
5. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP, pp. 373–376 (1996)
6. Sawant, R., Virani, H.G., Desai, C.: Database selection for Concatenative speech synthesis With novel endpoint detection algorithm. IJAIEM **2**(5), 173–180 (2013)
7. Psutka, J.: Communication with Computer by Speech (in Czech). Academia, Prague (1995)

8. Radová, V.: UWB S01 corpus - a Czech read-speech corpus. In: Proceedings of ICSLP2000, vol. IV, pp. 732–735. Beijing (2000)
9. Anguera, X., Luque, J., Gracia, C.: Audio-to-text alignment for speech recognition with very limited resources. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
10. Sproat, R., et al.: Normalization of non-standard words. Comput. Speech Lang. **15**(3), 287–333 (2001)
11. Musaev, M., Khujayorov, I., Ochilov, M.: The use of neural networks to improve the recognition accuracy of explosive and unvoiced phonemes in Uzbek language. Inf. Commun. Technol. Conf. **2020**, 231–234 (2020). https://doi.org/10.1109/ICTC49638.2020.9123309
12. Ping, W., Peng, K., Chen, J.: ClariNet: parallel wave generation in end-to-end text-to-speech. In: Proc. ICLR (2019)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
14. Musaev, M., Khujayorov, I., Ochilov, M.: Automatic recognition of Uzbek speech based on integrated neural networks. In: Aliev, R.A., Yusupbekov, N.R., Kacprzyk, J., Pedrycz, W., Sadikoglu, F.M. (eds.) WCIS 2020. AISC, vol. 1323, pp. 215–223. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68004-6_28
15. Gonzalvo, X., Tazari, S., Chan, C.A., Becker, M., Gutkin, A., Silen, H.: Recent advances in Google real-time HMM-driven unit selection synthesizer. In: Proc. Interspeech, pp. 2238–2242 (2016)
16. Ze, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proc. ICASSP, pp. 7962–7966 (2013)
17. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis. In: Proc. Interspeech (2017). arXiv:1703.10135
18. Oord, A.V.D., et al.: WaveNet: a generative model for raw audio (2016). arXiv:1609.03499
19. Musaev, M., Mussakhojayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., Atakan Varol, H.: USC: an open-source Uzbek speech corpus and initial speech recognition experiments. In: Karpov, A., Potapova, R. (eds.) SPECOM 2021. LNCS (LNAI), vol. 12997, pp. 437–447. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_40
20. Abdullaeva, M., Khujayorov, I., Ochilov, M.: Formant set as a main parameter for recognizing vowels of the Uzbek language. Int. Conf. Inf. Sci. Commun. Technol. **2021**, 1–5 (2021). https://doi.org/10.1109/ICISCT52966.2021.9670268
21. Raximov, R., Primova, H., Ruziyeva, Z.: Methods of recognizing texts in different images. In: International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities (2021). http://www.icisct2021.org/
22. Fazliddinovich, R.M., Abdumurodovich, B.U.: Parallel processing capabilities in the process of speech recognition. Int. Conf. Inf. Sci. Commun. Technol. **2017**, 1–3 (2017). https://doi.org/10.1109/ICISCT.2017.8188585
23. Musaev, M., Rakhimov, M.: A method of mapping a block of main memory to cache in parallel processing of the speech signal. Int. Conf. Inf. Sci. Commun. Technol. **2019**, 1–4 (2019). https://doi.org/10.1109/ICISCT47635.2019.9011946
24. Rakhimov, M., Ochilov, M.: Distribution of operations in heterogeneous computing systems for processing speech signals. In: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–4 (2021). https://doi.org/10.1109/AICT52784.2021.9620451