SATBAYEV UNIVERSITY

Computing & Engineering

Volume 1 (2023), Issue 3, 29-35

https://doi.org/10.51301/ce.2023.i3.06

Study of the transformation of Kazakh language speech into text data

A. Kursabayeva[1,2*]

[1]*Satbayev University, Almaty, Kazakhstan*
[2]*Suleyman Demirel University, Almaty, Kazakhstan*

**Corresponding author: aakursabayeva@gmail.com*

Abstract. This article investigates the use of VOSK-based voice recognition model for the Kazakh language. In particular, it provides a comparative analysis between two variants of VOSK speech recognition models: VOSK big and VOSK small. The assessment is carried out within the framework of the Kazakh language using the KazakhTTS dataset, prepared in 2021 by the ISSAI team. The results of the experiment, presented in the form of a Word Error Rate (WER), showed that VOSK big shows a better result (51%) compared to VOSK small (55%). However, it was pointed out that there are limitations in the recognition of word endings and that some errors occur in speech recognition. A discussion of the results highlights the potential of the model and points to the need for further refinement and training on more diverse data. The key conclusions are outlined in the conclusion, along with potential directions for further study in the area of Kazakh speech recognition.

Keywords: *speech recognition, Kazakh language, VOSK, audio.*

1. Introduction

In today's world, where the flow of information is constantly increasing and the importance of using information quickly is incredibly high, speech-to-text translation is becoming very important. This process allows for the efficient conversion of spoken audio and video material into a convenient and accessible text format. With the development of modern artificial intelligence technologies, the automatic transcription of speech into text has become an important part of our digital lives [1]. It is essential for many applications, from on-device speech recognition systems to the creation of transcriptions from video and audio recordings. In the speech-to-text process, language idiosyncrasies and dialects can pose significant challenges to algorithms and systems. The study of the conversion of Kazakh spoken language into textual data is very important in today's world, where the processing of large amounts of data and the development of artificial intelligence are becoming increasingly important [2]. This process has several important aspects and perspectives that make it important for our society and for the future.

First, such research is very important in the context of the development of speech technology and artificial intelligence. The conversion of spoken language into text is essential for developing automatic speech recognition (ASR), machine translation and other language applications. These technologies are very important for modern communication, education and business.

Secondly, this method's observation is significant for the maintenance and improvement of cultural and linguistic heritage. The Kazakh language is an essential a part of Kazakh culture, and its maintenance and merchandising withinside the virtual international play an essential role. The conversion of spoken language into textual content makes it feasible to create virtual archives, instructional substances and different assets in Kazakh.

The Kazakh language has its personal grammatical structure, which differs from English and different European languages. This consists of the declension and conjugation of nouns and verbs, in addition to unique phrase order rules. Properly spotting and deciphering the grammatical functions of Kazakh is a mission for speech processing technologies. The Kazakh language has its own grammatical features, which pose a challenge to speech processing technologies. For example, Kazakh uses a system of declension and conjugation of nouns and verbs. This means that the forms of words can change depending on their role in the sentence and other contextual factors. Consider the following example: «Әдемі» (beautiful) it may vary depending on its role in the sentence: «әдемі», «әдемінің», «әдеміге», «әдеміге» and so on [3].

These changes present a challenge for speech recognition algorithms, as they must consider words in their basic form and modified forms in various contexts. Another aspect is the use of the Cyrillic alphabet, which contains additional letters that are not present in the English alphabet. For example, «ғ» and «қ» represent sounds that are not present in English and require special recognition. «і» and "ә" also represent unique graphemes that do not exist in English. Such features create challenges for speech processing technologies, since algorithms must be configured to recognize and correctly interpret these unique elements in the process of converting oral speech in Kazakh into text data.

Thirdly, this observe has realistic programs in education, fitness care, regulation enforcement and plenty of different fields. Textual facts from spoken language may be used to create transcriptions of lectures, clinical records, courtroom docket transcripts, and plenty of different documents, which enables entry to statistics and improves methods in those fields.

All those complexities make the undertaking of reworking Kazakh speech into textual facts a mission for researchers and

engineers withinside the area of speech processing and synthetic intelligence. The main aim of this article is to investigate a technique for converting spoken Kazakh into written data while maintaining accuracy and usability.

The novelty of this observation lies withinside the exploration of the transformation method of spoken Kazakh language into textual facts, which entails the usage of superior herbal language processing strategies and linguistic analysis.

Text facts generated from spoken Kazakh may be hired to decorate accessibility for people with listening to impairments. It may be used to increase subtitles and closed captions for motion pictures and different multimedia content. This generation has the capacity to gain an extensive variety of sectors and enhance the lives of Kazakh-talking people and communities.

In the studies of this dissertation, three main objectives are set:

· Assessment of the precision of Kazakh voice recognition: The evaluation of contemporary technologies' ability to translate uttered Kazakh words into text involves an analysis of both the mistake rate and general recognition skills.

· Applications to be studied include the possible effects of improved Kazakh voice recognition on AI language integration, machine translation, and other critical domains requiring precise speech processing.

· Research questions:

· What is the accuracy and efficiency of existing methods for transforming Kazakh speech into text data, and how do they cope with the phonetic and grammatical features of the Kazakh language?

· What use cases are most in demand for Kazakh speech-to-text technology, and what are the benefits and limitations associated with these scenarios?

1.1. Literature review

The era of speech popularity has undergone considerable changes in recent years, with a developing emphasis on addressing the demanding situations posed with the aid of using the variety in spoken language.

Phonetic parameters for speech recognition. Hallet and Stevens present a speech recognition model and software that is based on an analysis-by-synthesis approach [4]. The approach enables the transcription of spoken language into written form by extracting phoneme sequences from time-varying input spectra. To efficiently recognize speech, the system makes use of generative principles; interestingly, the same calculations are applied to both speech production and identification. To determine phonetic parameters, the model emphasizes the importance of generative rules rather than direct speech structure activation.

The significance of managing variations in natural speech. Benzeghiba et al. recognize the major gains achieved in spoken language systems and automated speech recognition (ASR) [5], while emphasizing persistent technological challenges. Environmental factors such as background noise can exacerbate sensitivity, and the amount of grammatical and semantic knowledge that can be expressed is limited. For applications like directory assistance with huge active vocabularies and in the presence of foreign accents, the authors emphasize the need of handling true speech variation.

Multidimensional realization of speech. In addition to geographical variances and speaker characteristics, the study recognizes the intricate relationships between sociolinguistics, environment, gender, speaking tempo, accent, and style. These characteristics greatly increase the complexity of the modeling process, particularly in situations where training resources are few. The authors look at methods for improving ASR modeling and analysis, emphasizing how to make the system more resistant to speech fluctuation.

Novel methods for the acknowledgment procedure. To provide effective text or speech processing for human-machine communication, the Neha Chadha emphasizes the need for developing a recognition system. They include all the current speech recognition techniques and algorithms along with their advantages and disadvantages [6]. Based on the evaluation, there is room for innovation in coming up with new ideas for the recognition process that might produce better results than existing methods. Ambient noise issues, which reduce the effectiveness of audio input devices, are among the challenges discovered. Accuracy and reliability are further compromised by undesired input and subpar output results. The current systems do not have fault tolerance, and when users issue instructions before resources are prepared, it might impede their responsiveness and cause issues with synchronization across many apps.

Automatic speech recognition (ASR) and speech synthesis (TTS). Alexandre Trilla's research utilizes Natural Language Processing (NLP) approaches to two important fields of speech technology: text-to-speech synthesis (TTS) and automated speech recognition (ASR). This article gives an overview of the most recent advancements in NLP in many domains. The study emphasizes how important natural language processing (NLP) is to the examination of input text meant for voice conversion in the context of TTS synthesis [7]. The quality of the final speech is significantly influenced by how well the prior text-processing modules worked. Automatic speech recognition (ASR) is improved by natural language processing (NLP), assuming spoken input follows formal grammar rules. According to the paper, NLP can improve ASR capabilities by providing more natural interfaces with a certain degree of language competency. It examines context cues and N-gram language models to update language models to use linguistic knowledge. The research also helps with conversation systems and knowledge retrieval and provides methods for managing spontaneous speech through automated summarizing. It examines how ontologies are crucial for creating knowledge bases that support reasoning.

Autonomous speech recognition. Chiu et al. provide a simple and efficient method for improving autonomous speech recognition, particularly through self-supervised learning. Using a random projection quantizer, the method predicts speech signals that are masked and expressed as different labels. The model is taught to predict labels for the masked regions after speech signals have been masked during the pre-training phase. This approach shows promise by achieving word error rates (WERs) equivalent to state-of-the-art models for certain datasets [8]. In particular, the random projection quantizer simplifies the representation learning process, making self-supervised learning more efficient and palatable.

Model based on LSTM recurrent neural networks. In Jane Oruh's study, a deep learning model based on LSTM recurrent neural networks is shown to improve speech recognition. The model effectively handles continuous input streams without requiring significant bandwidth improvements by using an RNN as a "forget gate" in memory blocks [9]. The proposed model optimises the parameter usage of the standard LSTM architecture.

The paper compares the proposed LSTM-RNN model with CNN-based and sequential models using a widely known public benchmark dataset of spoken English numbers. The results show that the LSTM-RNN model outperforms other deep learning models with an accuracy rate of 99.36%.

Basics of Speech Popularity Systems. The aim of this taking a look at is to introduce the basics of speech popularity systems, consisting of their improvement and the latest upgrades that have been carried out to enhance their accuracy and robustness. Neha Jain et al. provide a radical evaluation of the mechanism, obstacles and techniques to overcome them [10]. In addition to reviewing the features that have taken place because of the introduction of traditional language popularity systems, this take a look at offers a quick evaluation of the variations among the fashions and algorithms that have been and are presently applied within the implementation of language popularity systems. Many frameworks and tools, together with the AURORA framework and the Voice Activity Detector, have been created to deal with these issues. The field of voice popularity era has superior to a degree in which matters now are substantially higher than they had been a few years ago. In the following few years, it's miles positive that robots will be able to hold close languages drastically higher than they might a few years ago.

Integrated approach and external factors. Anupam Choudhary and Ravi Kshirsagar give a thorough explanation of how artificial intelligence technology is used in the voice recognition process [11]. The source channel model, trigram model, class model, language model, and acoustic model are all described. It is stressed that voice recognition and natural language processing are terms used to describe artificial intelligence techniques for speaking with a computer in a natural language like English. The methodology takes a comprehensive voice recognition approach, considering contextual and external aspects as well as signal processing, acoustic modeling, and language modeling. The research also mentions how voice recognition accuracy can be greatly impacted by outside variables including ambient noise, location of the device, and environmental conditions.

In summary, these studies shed light on the dynamic field of speech recognition, where technological advances are made possible by natural language processing techniques. The techniques proposed, ranging from LSTM-RNN models to random projection quantization, represent viable ways to improve the accuracy and reliability of speech recognition systems.

Table 1. Literature review summary evaluation table

	Topic	Authors	Methodologies or key focus areas
1	Speech Recognition: A Model and a Program for Research	M. Hallet and K. Stevents,	Extracting phoneme sequences from input spectra via comparator comparison
2	Automatic speech recognition and speech variability: A review	M. Benzeghiba and et.	improving the accuracy of ASR analysis/modeling and resistance
3	Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey	Neha Chadha R.C. Gangwar, PhD Rajeev Bedi	recognition techniques and methods used in the current era with their pros and cons.
4	Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition	Alexandre Trilla	poverview of the latest advancements in Natural Language Processing (NLP) techniques
5	Self-Supervised Learning with Random-Projection Quantizer for Speech Recognition	Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, Yonghui Wu 1	model to anticipate concealed speech signals
6	Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition	Jane Oruh , Serestina Viriri , , Adekanmi Adegun	The main structure of LSTM consists of unique segments known as ‘‘memory blocks’’ in the hidden layer.
7	Speech Recognition systems- a comprehensive study of concept and mechanism	Neha Jain, Somya Rastogi	Extracting acoustic speech vectors from client queries, filtering, normalization, and signal segmentation are the main areas of attention in the speech conversion process.
8	Process Speech Recognition System using Artificial Intelligence Technique	Anupam Choudhary, Ravi Kshirsagar	comprehensive approach to speech recognition
9	An advanced NLP framework for high-quality Text-to-Speech synthesis	Catalin Ungurean, Dragos Burileanu	isolating the voice generating work from the application task and using SSML for data and control interaction
10	High-quality text-to-speech synthesis: an overview.	Thierry Dutoit	the potential applications and challenges of Text-To-Speech (TTS) synthesis systems.
11	Listen, Attend and Spell	William Chan	Listen, Attend and Spell (LAS) model
12	SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition	Daniel S. Park	SpecAugment model
13	Deep Speech: Scaling up end-to-end speech recognition	Awni Hannun , Carl Case	RNN model
14	Attention-Based Models for Speech Recognition	Jan Chorowski	recurrent neural network
15	fairseq S2T: Fast Speech-to-Text Modeling with fairseq	Changhan Wang	FAIRSEQ S2T model
16	Natural Language Processing for Text and Speech Processing: A Review Paper	Santosh Kumar Behera, Mitali M Nayak	a systematic and structured approach to collecting, categorizing, analyzing, and synthesizing literature
17	Speech Based Voice Recognition System for Natural Language Processing	Kavitha Raju	Extract voice features: frame, FFT, Mel, DCT.
18	Ims-speech: a speech to text tool	Pavel Denisov, Ngoc Thang Vu	Divide the recording into short speaking bursts. Transcribing each section requires performing ASR.
19	Towards Unsupervised Speech-to-Text Translation	Yu-An Chung, Wei-Hung Weng, Schrasing Tong, James Glass	emerging techniques in unsupervised speech processing and machine translation (MT).

20	Consecutive Decoding for Speech-to-text Translation	Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei LI	simplify the learning process for ST
21	Automatic speech recognition: a Comprehensive survey	PhD. Candidate Amarildo Rista; Prof. Dr. Arbana Kadriu	a systematic literature review on speech recognition approaches and techniques.
22	Speech to text conversion for multilingual languages	Yogita H. Ghadage, Sushama D. Shelke	During training, record speech utterances for each sentence, preprocess the speech signal. In testing, preprocess the speech utterance to be tested, segment into words, and extract features for each word.
23	Review of Algorithms and Applications in Speech Recognition System	Rashmi C R	Feature extraction: RCC, MFCC, LPC, LPCC, PLPC. Classification: VQ, HMM, GMM, SVM, MLP, DTW.
24	Real-Time Speech-To-Text /Text-To-Speech Converter With Automatic Text Summarizer using Natural Language Generation And Abstract Meaning Representation	K. P. Vijayakumar, Hemant Singh, Animesh Mohanty	complete replacement of a set of pipelines instructions with neural networks.
25	Voice recognition system: speech-to-text	Pranab Das, Vijay Prasad	utilization of Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) techniques. This system is composed of two primary components: Acoustic Signal Processing and Signal Interpretation.

2. Materials and methods

Within the framework of this study, the VOSK model was used to recognize Kazakh speech. During the experiment, careful preparation and preprocessing of the data played a key role, which provided a reliable basis for further analysis.

The experiments were performed using an extensive set of audio files, which made it possible to evaluate the versatility and adaptability of the VOSK model to various conditions and accents in the Kazakh speech. This was investigated by comparing the recognition results with pre-prepared text transcriptions. Therefore, in this experiment, we defined the goals as improving the accuracy of recognition and adapting the model to the peculiarities of Kazakh speech.

For experiment, author used part of a ready-made dataset called KazakhTTS. The KazakhTTS dataset was designed to provide comprehensive data collection on spoken Kazakh, starting with a thorough textual data collection process. The authors manually extracted more than 2,000 articles from various news sites, providing a wide range of topics including politics, business, sports and entertainment. The selected articles are then formatted in DOC format, considering the preferences of professional speakers regarding font size, line spacing and typeface for ease of reading. Segmentation and alignment of recorded audio files were performed by five native Kazakh translators. They used the Praat toolkit to divide the audio recordings into sentences and bring them in line with the corresponding text. The general statistics of the constructed dataset are presented in Table 2. The distribution of audio segment durations is shown in Figure 1.

Table 2. Dataset specifications

Category		Values
Duration	Total	10660 sec
	Mean	8.7 sec
	Max	22 sec
	Min	1 sec
Words	Total	3202650
	Mean	12.9
	Max	75
	Min	1

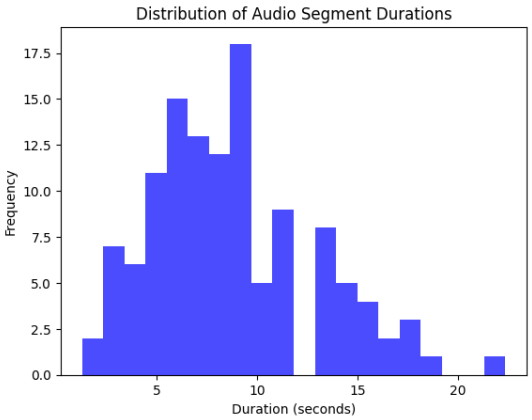

Figure 1. Distribution of audio segment duration

In research, author was able to effectively convert audio data into text format using the powerful tool VOSK. Based on cutting-edge deep learning technology, VOSK is a powerful open-source speech recognition tool that supports 17 languages and dialects for use as models at the time of writing. VOSK provides speech recognition for chatbots, smart home devices and virtual assistants. The model was developed by Alpha Cephei Mc. This library was created in 2019. VOSK is based on Kaldi speech recognition technology, but VOSK provides a simple interface for using this technology in various applications. The main developer of VOSK is Maxim Sergeev.

One of the reasons VOSK was chosen was because of the uncomplicated documentation. Another reason in favor of VOSK was that there is an explicit Kazakh model proposed to adapt the model. It is also capable of producing film subtitles and transcripts of lectures and interviews. VOSK's versatility as a research tool is largely due to its ability to handle a wide range of languages and dialects. In addition, its real-time functionality has allowed me to process large amounts of audio data quickly and efficiently, with immediate results.

VOSK is one of the open sources STT engines. VOSK must be installed before it can be used. There are two types of Kazakh VOSK model: big and small. For comparison, I used both options in my experiment. It can be installed via pip3. The VOSK model was downloaded and initialized prior to the start of the research, along with the installation of the required libraries, including VOSK, pydub and scikit-learn. The model was set up by selecting the version based on the requirements and objectives of the research.

Initializing the KaldiRecognizer recognizer with a model and sample rate indication was part of the processing

process. After that, the audio files were read in binary format, and the recognizer received the audio data from them. The identification outcome was stored for a future comparison with the real transcription taken from the relevant files.

Finding, sending, and analyzing audio data were among the processing procedures. The outcomes were then compared to accurate transcriptions. Each audio file underwent this procedure to assess the accuracy of recognition.

Therefore, analyzing the performance of a speech recognition model using the VOSK library provides a useful perspective on the use of automated speech recognition technologies.

The results can be used as a basis for further study and development in this area, providing new insights into audio processing.

3. Results and discussion

Table 3 shows the value of WER (the frequency of errors in words) and graph 2 and 3 shows a more detailed value. The word error rate, or WER, is a statistical indicator used to assess the accuracy of speech recognition. It calculates the percentage of errors in the recognized speech compared to the source text. The total number of inserts (incorrectly inserted words), substitutions (incorrectly recognized words) and deletions (missing words) divided by the total number of words in the source text is called WER and is indicated as a percentage.

Table 3. Table of word error rate values in percent

Model	WER (%)
VOSK Big	51
VOSK small	55

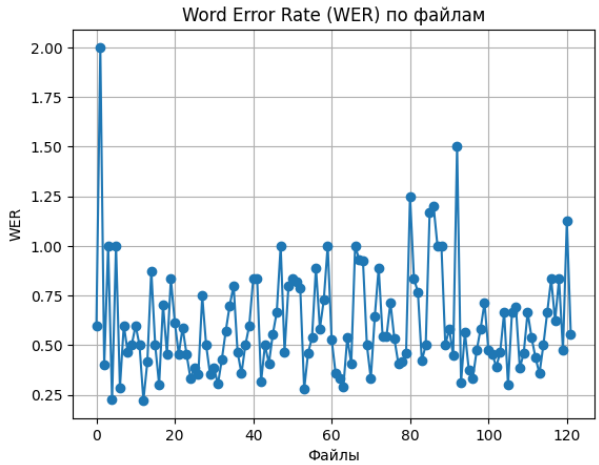

Figure 3. Word Error Rate by files for VOSK small model

These values indicate that the VOSK big model gives a better result (lower error rate) compared to the VOSK small model. During the comparative analysis, it was also noticed that VOSK big processes files a minute faster than VOSK small. The larger model recognizes a little better and takes up 40 times more space.

The results of the experiment demonstrated the shortcomings of the existing model's ability to identify word endings and certain instances of incorrect word recognition in Kazakh. In particular, there were problems with the accurate identification of endings, which in certain circumstances led to an incomplete experience of the final sounds.

These limitations could be due to variations in pronunciation, accents or linguistic features that weren't fully considered when training the model. It's also possible that some words were misidentified because the phonetic templates were difficult to match.

These findings highlight the importance of continuing to fine-tune and train the model on a larger and more diverse corpus of Kazakh language data in order to increase its accuracy and ability to recognize sentences with more complex structures. Despite these drawbacks, the model showed a degree of recognition that could be helpful in some situations but needs further fine-tuning for more accurate and comprehensive use.

The results presented in Table 3 and Figures 2 and 3 provide valuable information about the performance of the VOSK big and VOSK small models in the context of Kazakh speech recognition. The observed word error rate (WER) shows that the VOSK big model is superior to the VOSK small model with a lower error rate (51% compared to 55%). This indicates that the larger model has higher accuracy in recognizing Kazakh speech, which is consistent with expectations that larger models often produce improved results. The experiment revealed certain limitations in both models, especially in their ability to accurately identify the endings of words. There have also been cases of incorrect word recognition, possibly related to differences in pronunciation and linguistic nuances. These limitations emphasize the importance of considering phonetic variations and linguistic subtleties at the model learning stage.

4. Conclusions

In conclusion, although the experiment has demonstrated promising levels of recognition, there is room for improvement. However, it is important to note the disadvantages of the existing model, namely poor understanding of word endings and occasional speech recognition errors. These features emphasize the need for continuous development and improvement of the model through the use of an increasingly diverse training data. Continuous refinement and training on various data sets are important steps towards the development of a reliable speech recognition system for the Kazakh language. These achievements can have a great impact on many different areas, such as the creation of sophisticated translation tools and the introduction of Kazakh language skills into artificial intelligence systems.

Acknowledgements

Author would like to express sincere gratitude to supervisor, Dr. Dana Utebayeva, for her invaluable guidance, support and mentoring throughout the research process.

References

[1] Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A. & Turdalykyzy, T. (2019). Automatic Recognition of Kazakh Speech Using Deep Neural Networks. In: Nguyen, N., Gaol, F., Hong, TP., Trawiński, B. (eds) Intelligent Information and Database Systems. ACIIDS 2019. *Lecture Notes in Computer Science, Springer, Cham.* https://doi.org/10.1007/978-3-030-14802-7_40

[2] Orken, M., Dina, O., Keylan, A. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci Rep, (12), 8337*. https://doi.org/10.1038/s41598-022-12260-y
[3] Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Mukhsina, K., Keylan, A., BabaAli, B., Nabieva, G., Duisenbayeva, A. & Akhmetov, B. (2019). Continuous speech recognition of Kazakh language. *ITM Web of Conferences, (24), 01012*. https://doi.org/10.1051/itmconf/20192401012
[4] Halle, M. & Stevens, K. M. (1962). Speech recognition: a model and a program for research. *IRE Transactions on Information Theory, 8(2), 155–159*. https://doi.org/10.1109/TIT.1962.1057686
[5] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. & Wellekens, C. (2007, October). Automatic speech recognition and speech variability: *A review. Speech Communication, 49(10–11), 763–786*. https://doi.org/10.1016/j.specom.2007.02.006
[6] Chadha, N., Gangwar, R., & Bedi, R. (2015). Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey. *International Journal of Computer Applications, 131(11), 28–31*. https://doi.org/10.5120/ijca2015907471
[7] Trilla, A. (2009). Natural Language Processing Techniques in Text-to-speech Synthesis and Automatic Speech Recognition. *Universitat Ramon Llull, Barcelona, Spain*
[8] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, Yonghui Wu. (2022). Self-Supervised Learning with Random-Projection Quantizer for Speech Recognition. arXiv. https://doi.org/10.48550/arXiv.2202.01855
[9] Oruh, J., Viriri, S. & Adegun, A. (2022). Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. *IEEE Access, 10, 30069–30079*. https://doi.org/10.1109/access.2022.3159339
[10] Jain, N. & Rastogi, S. (2019). Speech recognition systems -a comprehensive study of concepts and mechanism. *Acta Informatica Malaysia, 3(1), 01–03*. https://doi.org/10.26480/aim.01.2019.01.03
[11] Anupam, C. & Ravi, K. (2012). Process Speech Recognition System using Artificial Intelligence Technique. International *Journal of Soft Computing and Engineering (IJSCE), 2(5)*
[12] Ungurean, C. & Burileanu, D. (2011). An advanced NLP framework for high-quality Text-to-Speech synthesis. *6th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Brasov, Romania*. https://doi.org/10.1109/SPED.2011.5940733
[13] Chan, W., Jaitly, N., Le, Q. V. & Vinyals, O. (2015). Listen, Attend and Spell. arXiv [Cs.CL]. *Retrieved from:* http://arxiv.org/abs/1508.01211
[14] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech*. https://doi.org/10.21437/interspeech.2019-2680
[15] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E. & Ng, A.Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. arXiv [Cs.CL]. *Retrieved from:* http://arxiv.org/abs/1412.5567
[16] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. arXiv [Cs.CL]. *Retrieved from*: http://arxiv.org/abs/1506.07503
[17] Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D., & Pino, J. (2022). fairseq S2T: Fast Speech-to-Text Modeling with fairseq. arXiv [Cs.CL]. *Retrieved from:* http://arxiv.org/abs/2010.05171
[18] Behera, S. K. & Nayak, M. M. (2020). Natural Language Processing for text and Speech Processing: a review paper. *Social Science Research Network. Retrieved from:* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3878634
[19] Raju, K. (2014). Speech based voice recognition system for natural language processing. *International Journal of Computer Science and Information Technologies, 5 (4), 5301-5305*
[20] Denisov, P. & Vu, N. T. (2019). IMS-Speech: A Speech to Text Tool. arXiv [Cs.CL]. *Retrieved from:* http://arxiv.org/abs/1908.04743
[21] Chung, Y.-A., Weng, W.-H., Tong, S., & Glass, J. (2018). Towards Unsupervised Speech-to-Text Translation. arXiv [Cs.CL]. *Retrieved from*: http://arxiv.org/abs/1811.01307
[22] Rista, A., & Kadriu, A. (2020). Automatic Speech Recognition: A Comprehensive survey. *SEEU Review, 15(2), 86–112*. https://doi.org/10.2478/seeur-2020-0019
[23] Mussakhojayeva, S., Khassanov, Y. & Varol, A. (2022). KazakhTTS2: Extending the Open-Source Kazakh TTS Corpus with More Data, Speakers, and Topics. *Accepted to LREC*
[24] Hansen, J.H.L. & Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine, 32(6), 74–99*. https://doi.org/10.1109/msp.2015.2462851
[25] Vijayakumar, K.P., Singh, H.K. & Mohanty, A. (2020). Real Time Speech to Text Text To Speech Converter With Automatic Text Summarizer using Natural Language Generation And Abstract Meaning Representation. *International Journal of Engineering and Advanced Technology, 9(4), 2361–2365*. https://doi.org/10.35940/ijeat.d7911.049420
[26] Prasad, V. (2015). Voice recognition system: speech-to-text. *Journal of Applied and Fundamental Sciences, 1(2), 191*. http://journals.dbuniversity.ac.in/ojs/index.php/JFAS/article/download/103/123

Қазақ тіліндегі сөйлеудің мәтіндік деректерге айналуын зерттеу

А. Курсабаева[1,2*]

[1]*Satbayev University, Алматы, Қазақстан*
[2]*Сүлейман Демирел Университеті, Алматы, Қазақстан*

**Корреспонденция үшін автор: aakursabayeva@gmail.com*

Аңдатпа. Бұл мақалада VOSK моделі арқылы қазақ тіліндегі сөйлеудің трансформациясы зерттеледі. Атап айтқанда, ол VOSK сөйлеуді тану моделінің екі нұсқасына салыстырмалы талдау жасайды: VOX big және VOSK small. Бағалау 2021 жылы ISSAI командасы дайындаған Kazakhs деректер жинағын пайдалана отырып, қазақ тілі шеңберінде жүргізіледі. Сөздегі қателік коэффициенті (WER) ретінде ұсынылған эксперимент нәтижелері VOSK big vosk small (55%) салыстырғанда жақсы нәтиже (51%) көрсететінін көрсетті. Алайда, сөздердің аяқталуын тануда шектеулер бар екендігі және сөйлеуді тану кезінде кейбір қателіктер болатындығы айтылды. Нәтижелерді талқылау модельдің әлеуетін көрсетеді және одан әрі әр түрлі мәліметтер бойынша нақтылау мен оқыту қажеттілігін көрсетеді. Қоры-

тындыда негізгі тұжырымдар, сондай-ақ қазақ тілін тану саласында одан әрі зерделеу үшін әлеуетті бағыттар баяндалған.

Негізгі сөздер: *сөйлеуді тану, қазақ тілі, VOSK, аудио.*

Исследование трансформации речи на казахском языке в текстовые данные

А. Курсабаева[1,2*]

[1]*Satbayev University, Алматы, Казахстан*
[2]*Университет Сулеймана Демиреля, Алматы, Казахстан*

**Автор для корреспонденции:* *aakursabayeva@gmail.com*

Аннотация. В этой статье исследуется трансформация речи на казахском языке на основе VOSK. В частности, в ней приводится сравнительный анализ двух вариантов модели распознавания речи VOSK: VOX big и VOSK small. Оценка проводится в рамках казахского языка с использованием набора данных KazakhTTS, подготовленного в 2021 году командой ISSAI. Результаты эксперимента, представленные в виде коэффициента ошибок в словах (WER), показали, что VOSK big показывает лучший результат (51%) по сравнению с VOSK small (55%). Однако было указано, что существуют ограничения в распознавании окончаний слов и что при распознавании речи возникают некоторые ошибки. Обсуждение результатов подчеркивает потенциал модели и указывает на необходимость дальнейшей доработки и обучения на более разнообразных данных. В заключении изложены ключевые выводы, а также потенциальные направления для дальнейшего изучения в области распознавания казахской речи.

Ключевые слова: *распознавание речи, казахский язык, VOSK, аудио.*

Received: 13 June 2023
Accepted: 15 September 2023
Available online: 30 September 2023