

Synthetic Data Generation for Kazakh Speech Separation and Diarization based on the use of neural networks

Dina Oralbekova^a, Orken Mamyrbayev^a, Larysa Azarova^b, Turdybek Kurmetkan^c,
Halyna Gordiichuk^d, Nurdaulet Zhumazhan^e, Daniel Sawicki^{*f}

^aInstitute of Information and Computational Technologies CS MSHE RK, Pushkin str. 125, 050010, Almaty, Kazakhstan

^bVinnitsia National Technical University, Khmelnytsky highway, 95, Vinnitsa, 21021, Ukraine,

^cAl-Farabi Kazakh National University, 71 Al-Farabi Avenue, 050040, Almaty, Kazakhstan

^dVinnitsia Mykhailo Kotsiubynskyi State Pedagogical University, 32 Ostrozhskogo str., 21001, Vinnitsa, Ukraine,

^eU.Joldasbekov Institute of Mechanics and Engineering, 28 Shevchenko str., 050010 Almaty, Kazakhstan,

^fLublin University of Technology, Nadbystrzycka 38D str., 20-618, Lublin, Poland

ABSTRACT

This paper explores the impact of various synthetic data generation methods on the performance of speech separation and diarization models. Three approaches are considered: simple audio track overlay, synthetic dialogue generation, and acoustic condition modeling. To evaluate their effectiveness, we used Conv-TasNet for speech separation and EEND-Conformer for diarization, both trained on a 400-hour Kazakh speech corpus. Experiments demonstrated that synthetic data can significantly enhance model performance when adapting to low-resource languages. The most effective method was synthetic dialogue generation, yielding results close to those obtained with real data for both speech separation and diarization. In contrast, acoustic condition modeling showed the highest deviations, indicating the need for further refinement. The findings confirm the potential of synthetic data for speech processing tasks. The proposed methods can improve the performance of automatic speech recognition models in scenarios with limited labeled data and challenging acoustic environments.

Keywords – synthetic data generation, speech separation, diarization, Kazakh language, Conv-TasNet, EEND-Conformer.

1 INTRODUCTION

Automatic speech recognition (ASR) plays a key role in modern sound data processing technologies and is widely used in voice assistants, machine translation, automatic subtitling, and human-computer interaction systems^{1,2}. An important task in the ASR field is speech signal separation and diarization, especially in multi-speaker environments such as video conferencing, transcription systems, and voice interfaces³.

Despite significant progress in ASR, the performance of models largely depends on the volume and quality of training data, particularly in cases of overlapping speech signals. Most modern solutions are developed for languages with abundant labeled data, whereas low-resource languages, including Kazakh, face a shortage of audio data. The limited amount of annotated recordings hinders model training, reducing their accuracy and adaptability to the specific phonetic and morphological characteristics of such languages.

One effective approach to addressing this issue is synthetic data generation, which allows for increasing the volume of training datasets, improving model robustness to various acoustic conditions, and enhancing their adaptation to real-world scenarios. This study examines three methods of synthetic data generation for speech separation and diarization: simple audio track overlaying, synthetic dialogue generation, and acoustic condition modeling.

The aim of this research is to evaluate the impact of different synthetic data generation methods on the quality of speech separation and diarization in the context of the low-resource Kazakh language.

*d.sawicki@pollub.pl

The novelty of this study lies in the first comparative analysis of three different synthetic data generation methods for Kazakh speech separation and diarization. The research assesses the influence of synthetic data on two state-of-the-art models: Conv-TasNet, designed for speech separation, and EEND-Conformer, used for diarization. The proposed approach is aimed at improving the quality of automatic speech separation and diarization for the Kazakh language and can be applied in various speech processing applications.

2 LITERATURE REVIEW

Mehta et al. (2024)⁴ investigated the use of synthetic data to improve model performance in audiovisual information processing tasks. The authors proposed a synthetic data generation method to compensate for the lack of annotated real data in resource-constrained conditions. Experiments showed that models trained with synthetic data achieved comparable or better results than those relying solely on real data. This study highlights the importance of synthetic data in enhancing model robustness and generalizability in real-world scenarios.

Hilmes et al. (2024)⁵ evaluated the impact of synthetic data on ASR system training. They used ASR training data to train a text-to-speech (TTS) system based on an architecture similar to FastSpeech-2. The model then generated synthetic versions of the original data, which were used to train ASR models exclusively. Three different ASR architectures were tested: an encoder-decoder with an attention mechanism, a hybrid deep neural network with a hidden Markov model (HMM-DNN), and a Gaussian mixture model with a hidden Markov model (GMM-HMM). The analysis revealed that different ASR models exhibit varying sensitivity to synthetic data. In particular, increasing the number of TTS model parameters and employing more complex speaker embeddings led to improved ASR accuracy.

Study⁶ introduced the DiffS4L pretraining framework, which leverages synthetic data to supplement a limited set of unlabeled speech data. The synthetic data is generated with varying levels of modifications using a diffusion model trained on unannotated recordings. The self-training model is first pretrained on a combination of real and synthetic data. Experimental results demonstrate that DiffS4L significantly improves ASR accuracy, reducing the word error rate (WER) of the pretrained HuBERT model by 6.26 percentage points in English ASR tasks.

Study⁷ presented a data augmentation method aimed at addressing the issue of limited data availability for online hate speech detection in low-context environments. The authors generated synthetic training data based on a small set of hate speech examples in a high-resource language, such as English. They proposed three methods for creating new examples that preserve the hateful tone of the original text while altering the target of hostility. This approach was applied to generate training data for hate speech classification in Hindi and Vietnamese. Experiments showed that models trained on synthetic data outperformed those trained only on the available in-domain examples.

Ghosh et al. (2024)⁸ introduced DARAG (Data- and Retrieval-Augmented Generative Error Correction)—a novel method designed to enhance generative error correction (GEC) in ASR systems both in-domain (ID) and out-of-domain (OOD). The authors found that standard GEC models struggle to generalize to unseen errors, particularly when tested on new data. This issue is even more pronounced when handling named entities (NE), as models often lack contextual information about them or encounter previously unseen NEs. To address these challenges, the authors enriched the GEC training set with synthetic examples generated using large language models (LLMs) and text-to-speech models. This approach helped the model learn to correct diverse errors, including domain-specific ones. DARAG achieved a relative WER reduction of 8%–30% in ID scenarios and 10%–33% in OOD scenarios.

Martius et al. ⁹ explored ways to enhance transcription quality in pretrained ASR models for maritime communication, focusing on accurately recognizing specialized terminology. Due to the lack of labeled maritime radio communication data, the authors created a synthetic training dataset adapted to regional maritime terminology. The generated synthetic speech was combined with human speech recordings and used to fine-tune an end-to-end ASR model under various conditions. A proprietary dataset was used for evaluation. Experimental results showed a significant improvement in model performance, with the proposed approach reducing WER by 13.46% compared to the pretrained baseline model, while recall for ship names and locations improved by 41.57% and 38.65%, respectively.

Study¹⁰ examined synthetic data generation methods for training siren identification models. The initial siren signals were obtained either from small datasets of recorded sounds available in public sources or through synthetic generation. The movement of the source, acoustic propagation, and Doppler effect were then simulated, and the resulting signal was mixed

with background noise. This process generated two synthetic datasets used to train three different convolutional neural networks, which were subsequently tested on real data not included in training. Results indicated that the proposed method, based on a combination of recorded signals and synthetic sound propagation modeling, provided the highest performance.

Liu et al. (2024)¹¹ introduced Libri2Vox, a new dataset that combines clean target speech from LibriTTS with interfering speech from VoxCeleb2, creating a large and diverse corpus with multiple speakers in realistic noisy conditions. To further enhance diversity, the authors included synthetic speakers generated using state-of-the-art speech synthesis models. Additionally, a curriculum learning strategy was implemented to enable target speech extraction (TSE) models to gradually adapt to increasingly complex conditions. Experiments with various TSE architectures demonstrated that the impact of the proposed approach varied depending on the model, with SpeakerBeam achieving the greatest improvement—an SDR increase of 1.39 dB on the Libri2Talker test set compared to the baseline model. These results confirm that diverse real data, synthetic speaker augmentation, and structured training strategies contribute to the development of more robust TSE systems.

Scheibler et al. (2022)¹² introduced DiffSep, a novel single-channel source separation method based on approximating stochastic differential equation (SDE) estimates. They developed a continuous-time diffusion mixing process that starts with separated sources and gradually transforms them into a Gaussian distribution centered on their mixture. This approach enables the application of generative modeling techniques based on probability estimation. The method consists of two key steps: training a neural network to approximate the probabilistic limit function or the diffusion mixing process, and solving the reverse SDE to iteratively separate sources from their mixture. Experiments with the WSJ0 2mix dataset confirmed the effectiveness of the proposed approach, demonstrating its potential for source separation tasks.

Sarkar et al. (2023)¹³ proposed a pretraining strategy and data augmentation technique based on multi-microphone renders from the synthetic dataset EnsembleSet and evaluated its impact on real chamber ensemble recordings from URMP. The use of synthetic data to supplement the training set significantly improved the performance of time-domain separation models, achieving up to a +5.14 dB gain when tested on real data. Furthermore, fine-tuning the models in combination with data augmentation yielded an improvement of up to +10.62 dB over the baseline in the chamber ensemble separation task. These findings confirm that synthetic data can be valuable in improving model performance in audiovisual information processing tasks.

3 METHODOLOGY

To create synthetic data, three methods were developed to model different scenarios of speech signal mixing.

3.1 Simple Overlay of Audio Tracks

This method is based on summing two audio recordings with a possible time shift. This technique allows for the creation of data that mimic natural voice overlaps, characteristic of multi-speaker dialogues and noisy environments. During experiments, audio files were randomly selected, and the time shift varied within a range of several hundred milliseconds to increase the variability of the synthetic data.

For implementation, Python tools and libraries were used, such as Librosa—for loading, processing, and analyzing audio files—and pydub—for handling audio formats and merging tracks.

The generation process consists of several steps (Fig. 1):

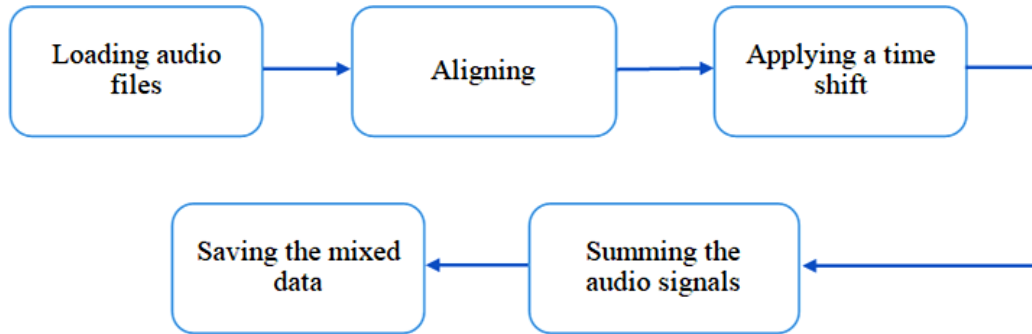


Figure 1. The method of summing two audio recordings

1. Loading two audio files and converting them into numerical arrays.
2. Aligning the duration of the audio files (if necessary) by trimming or padding with zeros.
3. Applying a time shift to the second audio file to create variation in the overlap.
4. Summing the audio signals with volume normalization to prevent overload.
5. Saving the resulting mixed audio file.

3.2 Synthetic Dialogue Generation

This method involves the sequential alternation of speech segments from different sources with pauses added between them. It allows for modeling scenarios typical of phone conversations, interviews, and conferences, where speakers take turns (Table 1). This approach is particularly useful for training models to recognize individual speaker turns, which is crucial for dialogue systems and multi-token transcription models.

The steps for generating synthetic dialogues are presented in Table 1.

Table 1. Steps for Generating Synthetic Dialogues

Step	Tasks	Applied Tools and Methods
1. Segmenting the original audio files	- Splitting audio into segments containing speech phrases - Detecting speech segments and removing silence	- Voice Activity Detection (VAD)
2. Mixing and alternating segments	- Creating an alternating sequence of utterances - Inserting random pauses between utterances (0.2–1.5 sec) to simulate natural speech pauses	- Randomization of sequence - Random pause generation
3. Adding noise and background effects	- Inserting background noise (café, street, office) to simulate real-world environments	- torchaudio, audiomentations
4. Saving dialogues as new audio files	- Combining segments into a new audio file - Generating annotations with timestamped utterances	- pydub, soundfile

For segment extraction, modules such as webrtcvad and silero-vad can also be used.

3.3 Acoustic Condition Modeling

This method applies algorithms for simulating room acoustics (e.g., reverberation and echo modeling) to create realistic multi-channel recordings. This approach enables the generation of data that reflect the impact of environmental conditions

on audio signals, making trained models more robust to different recording conditions. For the experiments, room acoustics were simulated for various room sizes and configurations.

The steps for acoustic condition modeling are as follows:

1. Adding Reverberation (Room Effect)

To simulate reverberation, room impulse response (RIR) characteristics can be used. Pre-existing RIR datasets, such as OpenAIR, can be utilized. The convolution of an audio file with an impulse response is performed using `scipy.signal.fftconvolve`.

2. Adding Background Noise

Adding background noise improves the model's robustness to challenging acoustic conditions. Various noise datasets can be used, such as MUSAN dataset (music, noise, speech), DEMAND dataset (various noise conditions) and UrbanSound8K (urban environment noises). Noise is added to the audio file with different signal-to-noise ratio (SNR) levels using `pydub`.

3. Echo Effect and Multi-Channel Recording

Adding echo can be beneficial for testing models in complex conditions. Echo simulation with a delay (e.g., 50–300 ms) can be performed using `scipy.signal.lfilter`. Additionally, multi-channel audio can be recorded using multiple microphones.

EXPERIMENTS

Data. For the experiments, a 400-hour Kazakh speech corpus was used, containing audio files with transcripts. This corpus, collected by the Laboratory of Computer Engineering for Intelligent Systems, included recordings from Kazakh speakers of different ages, genders, and dialects. The audio data covered both studio and real-world noisy conditions, while the transcriptions were annotated considering the morphological features of the Kazakh language. This allowed for a more precise analysis of model errors and the consideration of linguistic characteristics in data processing.

The primary goal of this study is speech separation, focusing on cases where multiple speakers talk simultaneously. To address this, we developed and tested three synthetic data generation methods aimed at improving speech separation quality.

When overlaying audio tracks of 2–3 speakers with a time shift, the dataset size increased by 1.5 times. When modeling dialogues by combining utterances from different speakers, the dataset size grew by 1.2 times. When simulating acoustic conditions, the amount of data doubled.

Model Configuration. For the experiments, two different architectures were selected for different purposes: End-to-End Neural Diarization (EEND)¹⁴ and Conv-TasNet¹⁵.

Conv-TasNet (Convolutional Time-domain Audio Separation Network) is designed for separating multiple speakers' voices from a mixed audio signal. Unlike traditional methods operating in the frequency domain, this architecture works directly in the time domain, offering advantages in processing signals with high accuracy and low latency. The model consists of three key components: 1) Encoder – converts the input signal into time-domain representations. 2) Separation Network – learns masks to separate different sources and 3) Decoder – reconstructs the extracted signals back into the original audio format. A detailed description of these components is provided in¹⁵.

The EEND (End-to-End Neural Diarization) method automatically determines who is speaking and when, without requiring stepwise segmentation. Unlike traditional diarization approaches such as x-vector + clustering, EEND utilizes multi-label prediction and a self-attention mechanism, making it particularly effective in multi-speaker scenarios. Like Conv-TasNet, EEND consists of three elements: an encoder that converts the acoustic signal into embeddings, a self-attention mechanism, and a decoder that models dependencies between speech segments and predicts the probability of each speaker's presence at each time frame¹⁶. For the experiments, EEND was implemented using a Conformer-based model, with the full architecture described in¹⁴. A detailed explanation of the architecture and potential model modifications is beyond the scope of this paper, as the primary focus is on synthetic data generation and model adaptation to the Kazakh language.

Conv-TasNet operates in the time domain and employs convolutional blocks to extract individual speech sources. During the experiments, the encoder kernel size was set to 15 ms, with 512 filters in the encoder. The hidden layer size in the separation module was 256, and the number of convolutional blocks was 8. The receptive field length was set to 200 ms. Adam was used as the optimizer with a learning rate of $1e-3$, which was reduced upon performance plateauing. The batch size was 16, and training was conducted for 100 epochs.

For the EEND-Conformer model, the embedding size was set to 512, and the number of Conformer layers was 12. The self-attention mechanism used 6 attention heads, and the hidden layer size was 512. Binary Cross-Entropy (BCE) was used as the loss function. Dropout was set to 0.3 for regularization. AdamW was used as the optimizer, as it regulates weights more effectively and improves model generalization. The learning rate was set to $1e-4$. Due to high memory consumption, the batch size was reduced to 8, and training lasted 100 epochs.

ANALYSIS OF RESULTS

Experiments were conducted to evaluate Conv-TasNet and EEND-Conformer on three synthetic data generation methods as well as real data.

Conv-TasNet was assessed using SDR (dB) and SI-SNR (dB) metrics, with the results presented in Table 2.

Table 2. Evaluation Results of Conv-TasNet on Real and Synthetic Data

Data Method	SDR (dB)	SI-SNR (dB)
Real data	12.5	11.8
Simple audio track overlay	10.2	9.5
Synthetic dialogue generation	12.1	11.3
Acoustic condition simulation	9.8	9.0

From the table, it is evident that Conv-TasNet achieved the best results on real data, which is expected since synthetic data may not fully capture the acoustic characteristics of real recordings. Among the synthetic data methods, synthetic dialogue generation yielded the best results, as its SDR and SI-SNR values are closest to those of real data.

For EEND-Conformer, the DER (%) metric was used, and the results are presented in Table 3.

Table 3. Evaluation Results of EEND-Conformer on Real and Synthetic Data

Data Method	DER (%)
Real data	12.3
Simple audio track overlay	18.5
Synthetic dialogue generation	13.1
Acoustic condition simulation	20.1

The DER value indicates how accurately the model identifies speakers. A lower DER corresponds to better diarization performance. As seen in the table, the model performed best on real data (12.3%), while among synthetic methods, synthetic dialogue generation had the lowest error rate (13.1%), indicating that it better approximates real-world conditions. Conv-TasNet demonstrated high-quality speech separation on real data, achieving SDR = 12.5 dB and SI-SNR = 11.8 dB. Among synthetic methods, synthetic dialogue generation produced results closest to real data (SDR = 11.3 dB), whereas acoustic condition modeling had the worst performance (SDR = 9.8 dB), likely due to lower-quality synthetic data.

For EEND-Conformer, the DER on real data was 12.3%, confirming the model's ability to accurately identify speakers. Among synthetic data methods, synthetic dialogue generation again performed best (DER = 13.1%), while acoustic condition modeling had the highest error rate (DER = 20.1%), indicating significant divergence from real-world data.

Thus, synthetic dialogue generation proved to be the most effective method, yielding the best results for both models, whereas the other two methods require further refinement due to their lower quality compared to real data. While synthetic data can be used, achieving high accuracy requires adapting them to the characteristics of real audio recordings.

CONCLUSION

This study analyzed three synthetic data generation methods for speech source separation and diarization: simple audio track overlay, synthetic dialogue generation, and acoustic condition modeling. Their impact was evaluated using two models: Conv-TasNet for speech separation and EEND-Conformer for diarization. The results demonstrated that the quality of synthetic data significantly influences model performance. The most effective method was synthetic dialogue generation,

which produced results closest to real data for both Conv-TasNet (SDR = 11.3 dB) and EEND-Conformer (DER = 13.1%). This suggests that synthetic dialogue generation can create training data that closely resemble real-world conditions. Thus, the study confirms that synthetic dialogue generation is the most promising approach for generating artificial training data. However, further refinement of acoustic condition modeling and simple audio track overlay is necessary to improve their alignment with real recordings.

Acknowledgments. *This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19174298).*

REFERENCES

- [1] Scharenborg, O., Ciannella, F., Palaskar, S., Black, A.W., Metze, F., Ondel, L., Hasegawa-Johnson, M.A. Building an ASR System for a Low-resource Language Through the Adaptation of a High-resource Language ASR System: Preliminary Results, (2017).
- [2] Oralbekova, D., Mamyrbayev, O., Othman, M., Alimhan, K., Zhumazhanov, B., Nuranbayeva, B., Development of CRF and CTC Based End-to-End Kazakh Speech Recognition System," in Intelligent Information and Database Systems. ACIIDS 2022, N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, and E. Szczerbicki, Eds., Lecture Notes in Computer Science, vol. 13757. Cham: Springer, 2022, pp. 497–508. doi: 10.1007/978-3-031-21743-2_41
- [3] Oralbekova, D., Mamyrbayev, O., Kurmetkan, T., Zhumazhan, N., "Recent Methods and Algorithms in Speech Segmentation Tasks," in Advances in Computational Collective Intelligence. ICCCI 2024, N. T. Nguyen et al., Eds., Communications in Computer and Information Science, vol. 2166. Cham: Springer, pp. 248–260 (2024). doi: 10.1007/978-3-031-70259-4_21
- [4] Mehta, S., Deichler, A., O'Regan, J., Moëll, B., Beskow, J., Henter, G.E., Alexanderson, S., Fake it to make it: Using synthetic data to remedy the data shortage in joint multi-modal speech-and-gesture synthesis, 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1952-1964 (2024).
- [5] Hilmes, B., Rossenbach, N., Schlüter, R., On the Effect of Purely Synthetic Training Data for Different Automatic Speech Recognition Architectures, Proc. 1st Workshop on Synthetic Data for Generative AI (SynData4GenAI), 46-50 (2024).
- [6] Gao H., Qian K., Ni J., Gan C., Hasegawa-Johnson M., Chang S., Zhang Y., "Speech self-supervised learning using diffusion model synthetic data", In Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 591, 14790–14810 (2024).
- [7] Khullar, A, Nkemelu, D., Nguyen, V.C., Best, M.L., "Hate Speech Detection in Limited Data Contexts Using Synthetic Data Generation", ACM J. Comput. Sustain. Soc. 2, 1, Article 4 (March 2024), 18 pages. (2024). <https://doi.org/10.1145/3625679>
- [8] Ghosh, S., Rasooli, M. S., Levit, M., Wang, P., Xue, J., Manocha, D., Li J. "Failing Forward: Improving Generative Error Correction for ASR with Synthetic Data and Retrieval Augmentation" (2024). 10.48550/arXiv.2410.13198.
- [9] Martius, C., Nakilcioglu, E.C., Reimann, M., & John, "Refining maritime Automatic Speech Recognition by leveraging synthetic speech", Maritime Transport Research (2024).
- [10] Damiano, S, Cramer, B, Guntoro, A. and van Waterschoot, T., "Synthetic data generation techniques for training deep acoustic siren identification networks" Front. Sig. Proc. 4:1358532. (2024). doi: 10.3389/frsip.2024.1358532
- [11] Liu, Yun & Liu, Xuechen & Miao, Xiaoxiao & Yamagishi, Junichi. "Libri2Vox Dataset: Target Speaker Extraction with Diverse Speaker Conditions and Synthetic Data" (2024). 10.48550/arXiv.2412.12512
- [12] Scheibler, R., Ji Y., Chung, S. -W., Byun, J., Choe, S. and Choi, M. -S., "Diffusion-Based Generative Speech Source Separation," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 1-5, (2023). doi: 10.1109/ICASSP49357.2023.10095310

- [13] Sarkar, S., Thorpe, L., Benetos, E. and Sandler, M., "Leveraging Synthetic Data for Improving Chamber Ensemble Separation," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 1-5 (2023). doi: 10.1109/WASPAA58266.2023.10248118
- [14] Liu, Y., Han, E., Lee, C., & Stolcke, A. "End-to-end Neural Diarization: From Transformer to Conformer". ArXiv, abs/2106.07167 (2021).
- [15] Luo, Y. and Mesgaran, N., "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256-1266 (Aug. 2019). doi: 10.1109/TASLP.2019.2915167.
- [16] Mamyrbayev, O., Wojcik, W., Titova, N., Pavlov, S., Oralbekova, D., Aitkazina, A., & Zhumazhan, N. Development of a thermodynamic model for optimization of processes in crop production. Eastern-European Journal of Enterprise Technologies, 6(8 (126), 25–34 (2023). <https://doi.org/10.15587/1729-4061.2023.290294>.