MINISTRY OF HIGHER EDUCATION, SCIENCE AND INNOVATION OF THE REPUBLIC OF UZBEKISTAN

TASHKENT STATE UNIVERSITY OF ECONOMICS

DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE: PROBLEMS, INNOVATIONS AND TRENDS

1st International Scientific - Practical Conference

CONFERENCE PROCEEDINGS

SEPTEMBER 11, TASHKENT 2024

ORGANIZING COMMITTEE

CHAIRMAN AND DEPUTIES

Tulkin Teshabayev

Rector of Tashkent State University of Economics

Gulnora Abduraxmanova

Vice-rector for scientific affairs and innovations

Sultonali Mehmonov

Vice-rector for academic affairs

Komila Karimova

First rector for youth issues and spiritual and educational affairs

Sherzod Sindarov

Vice-rector for international cooperation

Ulug‘bek Xalikov

Vice-rector for international cooperation

Nodir Akbarov

Dean of the Faculty of Digital Economy

Sanjar Mirzaliyev

Head of the Department of Scientific Research and Innovation

MEMBERS OF THE SCIENTIFIC AND TECHNICAL COMMITTEE

Bahodir Muminov

Head of the Department of Artificial Intelligence, Professor

Dilshod Mirzaev

Head of the Department of Information Systems and Technologies

Diyora Khashimova

Faculty of Digital economy, deputy dean

Rashid Nasimov

Associate professor of the Department of Artificial Intelligence

Sharofutdin Xoshimxodjaev

Associate professor of the Department of Artificial Intelligence

Guzal Belalova

Associate professor of the Department of Artificial Intelligence

Dilmurod Mirzaaxmedov

Senior teacher of the Department of Artificial Intelligence

Sanjar Muhammadiev

Senior teacher of the Department of Artificial Intelligence

Muminbek Khayrullayev

Assistant of the Department of Artificial Intelligence

Mamur Shuhratov

Assistant of the Department of Artificial Intelligence

Elyor Egamberdiev

Assistant of the Department of Artificial Intelligence

Ziyoda Norqulova

Assistant of the Department of Artificial Intelligence

FAST VOICE FILTERING IN A FEW STEPS USING VOICE CONVERSION AS A POST-PROCESSING MODULE ADAPTATION OF A SPEAKER FROM UZBEK TEXT TO SPEECH

Azamat Kakhorov,
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,
Uzbekistan,
azamat@tuit.uz

Mamur Shukhratov,
Tashkent State Economic University,
Uzbekistan.

A B S T R A C T **Text-to-Speech (TTS) systems developed in recent years require hours of recorded speech data to generate high-fidelity human-like synthetic speech. Low resources or small amount of speech can lead to several problems in the development of TTs models, which makes it difficult to train TTS systems with limited resources. This paper proposes a new low-resource TTS method called Voice Filter that uses only one minute of the target speaker's speech. It applies Voice Conversion (VC) as a post-processing module added to an already existing high-quality TTS system, which marks a conceptual change in the current TTS paradigm by recasting the multi-frame TTS problem as a VC task. In addition, it has been proposed to use a TTS system with controlled duration to create a parallel speech corpus that facilitates the VC task. The results show that Voice Filter outperforms modern multi-frame speech synthesis methods based on objective and subjective metrics using only one minute of speech from a diverse set of sounds, and at the same time with the Uzbek TTS model. remains competitive. 25 times more data.**

K E Y W O R D S Speaker Adaptation, Uzbek Text-To-Speech, Natural Language Processing, Few-Shot Learning, Voice Conversion

INTRODUCTION

Achievements in the field of artificial intelligence development today can truly surprise even those who are directly related to this field. One of the popular areas for the implementation of innovative projects has become the creation of digital applications compatible with computers and mobile devices, capable of analyzing and interpreting incoming data presented in voice or text form, without the use of special linguistics, template commands and formulations - that is, in the usual human construction of phrases and word combinations. How do algorithms and models of the natural language processing system work, what technologies, methods and tools are used for machine learning NLP (Natural Language Processing), and what does this give in terms of further development of the area in question?

Although modern text-to-speech (TTS) technologies are capable of producing high-quality synthetic speech in various scenarios, sometimes insufficient results can be achieved. Achieving very high-quality TTS usually requires several hours of studio-quality data from one or more speakers [1, 2]. Therefore, reducing the amount of speech data to several hours limits the quality and intelligibility of these systems. Since it is not always possible to collect several hours of speech data, especially when scaling TTS voices to a large number of new speakers, the problem of generating TTS voices with limited resources has been widely studied [2, 3,]. Many studies are also going in this direction. Because the issue of resources is always relevant. Of course, to build high-quality TTS systems in scenarios where, for example, the target speaker's speech is only three minutes long, we first aim to capture the identity of the existing speaker, and phonetically and prosaically we need to delay the modeling of variations. Given the lack of resources and the impossibility of replenishing them, the work will be ideal. Speaker identification can outperform TTS systems. Given the smallness of the given speech, the speaker is based on adaptation, in which the parameters of the multi-speaker model are optimized over many samples by repeatedly retraining the target speaker [5, 6].

As the field of text-to-speech technology advances, researchers are exploring innovative methods to overcome the challenges posed by limited data. One promising area of development is the use of deep learning models that can generalize well from small datasets, leveraging techniques such as transfer learning and data augmentation to enhance model robustness. Additionally, there is growing interest in cross-lingual TTS systems that can utilize data from multiple languages to improve speech synthesis quality, even for low-resource languages. These approaches not only enhance the accessibility and inclusivity of TTS technologies but also pave the way for more personalized and context-aware voice applications. By focusing on these strategies, the future of TTS could see widespread adoption across various industries, from virtual assistants and customer service bots to educational tools and accessibility solutions for individuals with disabilities [7].

The adaptation process focuses on altering only the speaker's identity, which comprises the speech attributes that define the target speaker as an individual. To control speaker identity during multi-step adaptation, methods such as vector quantization models [9], U-Net architectures [8, 9], attention mechanisms [10], or a combination of loss functions [18, 19] are employed.

Typically, speaker identification can be performed using images that have been passed through other external speaker verification systems or trained together with a base model [12, 20, 21]. In the adaptation process, studies have proposed optimizing all model parameters [11], selected components [13, 14, 15], or focusing on external aspects of the speaker instead [11]. An alternative approach is to solve the matching problem using data augmentation. This can be achieved using traditional signal processing methods, but more sophisticated

methods propose generating high-quality synthetic data for the target speaker using a voice transformation (VC) model [9]. The TTS system is then re-optimized by combining the natural and synthetic data.

However, there are limitations associated with these approaches. Wang et al. [18] argue that when using speaker adaptation strategies, a single architecture is tasked with modeling both linguistic content and speaker identity. Since it is not entirely clear which model parameters govern the speaker's identity within this framework, the effectiveness of parameter adaptation may be diminished. Moreover, fine-tuning a complex architecture on a limited number of samples can easily result in overfitting, thereby reducing overall quality and clarity. On the other hand, data augmentation techniques still require at least 15 minutes of training data from the target speaker to optimize TTS models, making them unsuitable for scenarios with extremely limited resources.

In this paper, we tackle the challenge of ultra-low-resource TTS by employing VC as a post-processing module, which we refer to as the "Voice Filter," applied on top of a high-quality single-speaker TTS model. This single-speaker TTS model is also utilized to generate a synthetic parallel corpus for training the Voice Filter. Our approach offers the following innovations and benefits:

(1) The entire process is made modular, separating it into a speech content generation task followed by a speaker identity generation task. This improves efficiency, reliability, and interpretability while allowing for task-specific adaptation;

(2) We exploit the advantages of parallel VC without requiring an available parallel corpus by synthetically creating speech pairs at the frame level using a TTS model with controlled duration.

In summarize, we have divided the challenge of creating a traditional Uzbek TTS voice into two tasks: speech content and speaker identity generation. This division allows us to minimize the amount of speech required to train a synthetic voice for a specific speaker identity to just one minute, thereby reducing the complexity of the problem. Consequently, the quality of the resulting synthetic speech matches the quality of TTS models trained on 25 times more data.

METHODS

Voice Filter addresses the challenge of Uzbek TTS voice generation with extremely limited resources by dividing the tasks of speech content and speaker identity generation, with Voice Filter focusing on the latter. This leads to greater modularization and more resilient speaker identity generation compared to the adaptation of the multi-speaker TTS model.

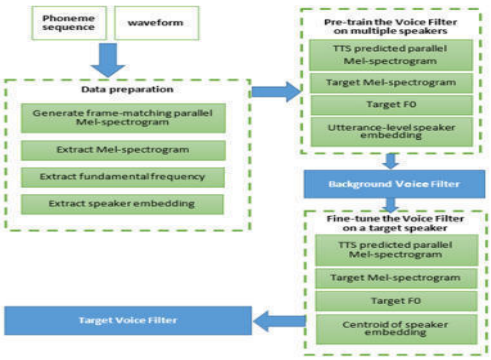

Figure 1. *Flowchart illustrating the preparation of synthetic parallel data and the training process for the proposed Voice Filter.*

Since Voice Filter is solely responsible for speaker identity generation, it functions at a more fundamental level (Mel-spectrograms) than the entire TTS system (phonemes). We believe that this speech-to-speech task is simpler than the text-to-speech task, particularly in resource-constrained settings. The proposed VC (Voice Filter) module is placed between the single-speaker duration-controlled TTS model and the general-purpose neural vocoder. This setup allows us to produce Mel-spectrograms for any desired text using the TTS model, which is then assigned the corresponding speaker identity by Voice Filter and finally transformed into a time-domain waveform using the vocoder.

Until recently, achieving individual goals required the use of traditional machine learning algorithms, which entailed a rigid selection of architecture and manual feature engineering. The advent of neural networks has enhanced the quality of results, enabling the identification of general principles for working with NLP.

II.I. Generating a Synthetic Parallel Corpus

The proposed method begins with the synthetic creation of a parallel dataset for training the voice filter (Figure 1, data preparation). Utilizing a synthetic parallel case allows us to overcome the two major constraints of traditional parallel VC while retaining its advantages: (1) the need for a large parallel corpus of discourse between source and target speakers, which is challenging and costly to gather, is eliminated; and (2) the necessity to employ time-shifting methods to align the parallel corpus across multiple speakers is removed. This is feasible because the input data for the voice filter are not recordings but duration-controlled, structured synthetic speech. However, this approach requires two distinct datasets to generate the parallel corpus: a single-speaker corpus for the TTS system and a multi-speaker corpus for training the voice filter. The construction of the synthetic parallel case involves three steps:

Force-smooth all available phone-level data, which we achieved using the pre-trained Kaldi ASpIRE TDNN system, encompassing both single- and multi-speaker cases.

Train a duration-guided TTS system using the single-speaker corpus.

Generate transcripts and synthetic data corresponding to the phone-level duration of the baseline adaptive multi-speaker corpus using the trained duration-guided TTS system.

This three-step process yields a frame-level parallel corpus mapping between the synthetic single-speaker speech samples and the natural multi-speaker speech samples. It forms the training data for the proposed voice filter, which, to the best of our knowledge, represents a novel approach to the VC problem.

For the purposes of this paper and experiment, the single-talker corpus comprises 20 hours of high-quality speech data read by a male Uzbek speaker in a neutral speaking style. The multi-talker corpus consists of 120 gender-balanced male and female Uzbek speakers, with approximately 35 minutes of data per speaker, ensuring comprehensive phonetic coverage.

II.II. Model Training and Fine-Tuning

Training a Voice Filter model capable of generating speech from 1 minute of unseen speaker data involves a two-step process: (1) preliminary training of the model (Figure 1, VF

pre-tuning) and (2) fine-tuning on a minute of the target unseen speaker data (Figure 1, VF fine-tuning).

The preliminary Voice Filter is trained in a one-to-many fashion for 1 million steps using the entire synthetic parallel multi-speaker corpus created earlier. This model can transform into any of the speakers encountered during training but lacks the robustness to generalize to unseen speakers without further refinement.

The preliminary model is adapted to the target Voice Filter by fine-tuning all parameters for 1000 steps on one minute of the target speaker's speech in a one-to-one manner. The centroid of the target speaker embeddings at the utterance level is employed because, in our multi-shot scenario, fine-tuning on a consistent speaker embedding rather than variable utterance embeddings led to more stable models. We did not evaluate the impact on quality for non-target speakers after fine-tuning, but we assume that the resulting target Voice Filter is speaker-specific. Both the preliminary and target Voice Filter models are trained using the L1 spectral loss and the BAHODIR optimizer with standard settings.

II.III. Model Architecture

The Voice Filter model (Figure 2) processes 80-bin Mel spectrograms of equal length as input and output and is composed of 6 layers of size-preserving 1D convolutions with 512 channels and a kernel size of 5, incorporating batch normalization. This is followed by a unidirectional LSTM and a dense layer with 1024 nodes. We combine the target speaker embedding and the log-f0 contour with the hidden representation of the third convolutional layer. The speaker embedding is a 256-dimensional vector defined at the utterance level and passed to the frame level. The speaker verification framework utilized to derive the embeddings was trained on a dataset of multiple speakers and fine-tuned on the generalized end-to-end loss. We found that the log-f0 contour assists the model in better capturing the prosodic differences between the input and target speakers. As a result, Voice Filter does not need to learn how to modify the prosodic information between the source and target speakers but rather concentrates on the speaker-specific information. To extract the log-f0 contour from the target speech recordings, we employed the RAPT algorithm from the Speech Processing Toolkit (SPTK) with a threshold of 0 for distinguishing between voiced and unvoiced regions.

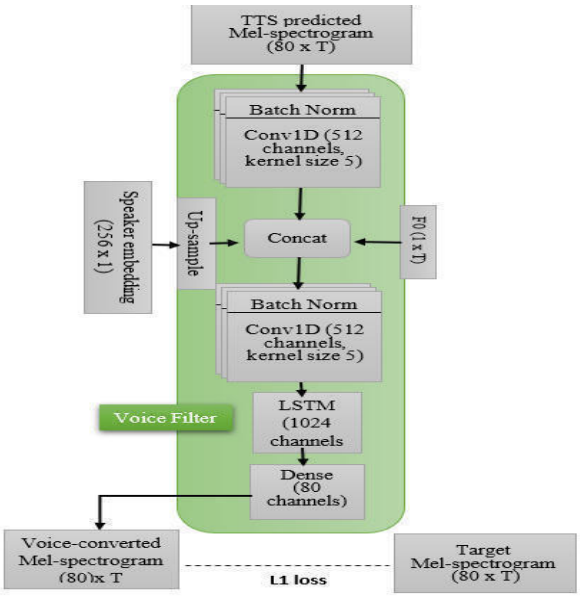

Figure 2. *Voice Filter Architecture (Proposed).*

II.IV. Model Inference

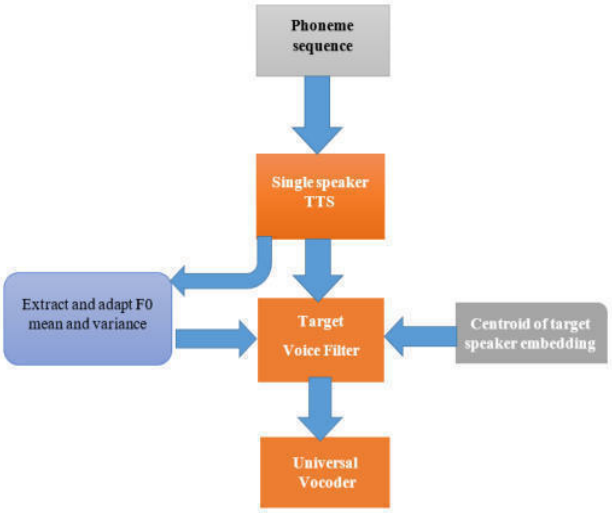

Figure 3. *Voice Filter inference flowchart (Proposed).*

The inference of the full model (Figure 3) requires us to sequentially run several models:

a. Estimating the f0 from the initial Mel-spectrogram and renormalizing it to match the mean and variance of the target speaker.

b. Generating the initial Mel-spectrogram for the desired text and predicted durations using the single-speaker TTS model.

c. Synthesizing the voice-converted Mel-spectrogram into a time-domain signal using a vocoder.

d. Transforming the initial Mel-spectrogram into the target speaker using the fine-tuned voice filter.

At this stage, we do not adapt the speech rate or phone duration to match the target speaker's characteristics, as these are difficult to estimate in extremely low-resource scenarios and can introduce significant artifacts.

EXPERIMENTAL SETUP

The models were assessed using both quantitative and perceptual metrics. For our evaluations, we selected 5 male and 5 female speakers, each with 60 test utterances, resulting in a total of 500 samples. Signal quality was quantitatively measured using the conditional Fréchet speech distance (cFSD) [22]. Specifically, a pre-trained XLSR-53 [22] wav2vec2.0 [24] model was employed to generate activation distributions for both the recordings and synthesized samples. These distributions were then compared using the Fréchet distance, providing a metric of how closely the generated speech matches the actual recordings. To quantitatively evaluate the speaker similarity metric, we used the average cosine distance between speaker embeddings (CSED) of the recordings and predicted samples.

The Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests were conducted to evaluate perceptions of naturalness, signal quality, talker similarity, and speaking style. Participants were presented with samples from the systems under evaluation side by side and asked to rate them

on a scale from 0 (worst) to 100 (best) for the metric being evaluated. We utilized the ClickWorker crowdsourcing platform to gather ratings for each test utterance from a group of 30 listeners. The target talker recordings were consistently included as a hidden top anchor system, and there was no requirement for at least one system to be rated as 100. Listeners were given a reference sample for both talker and style similarity ratings. The bottom anchor for talker similarity ratings was the voice-transformed samples of the most distant same-gender talker in the talker embedding space. The bottom anchor for style similarity ratings was the unfiltered TTS system. Paired two-tailed Student's t-tests with Holm-Bonferroni correction were used to assess the statistical significance of differences between the two systems at a threshold p-value of 0.1.

RESULTS

IV.I. Speech synthesis performance with extremely low resources

In this work, we compare the proposed system with two state-of-the-art multi-talker technologies that have demonstrated strong performance in low-resource speech synthesis: the CopyCat (CC) model [25] with the f0 padding condition (preliminary results indicated that the CopyCat model with the f0 padding condition exhibited better signal quality stability and provided a fairer comparison) and the Multi-Talker Controlled Duration TTS (MS-TTS) model [25] without the data padding component. Both models were trained on the same dataset and under the same conditions described in *Section 2.2*, including fine-tuning with 1000 steps per 1 minute of target speaker data. The objective metrics and MUSHRA perceptual scores are presented in Tables 1 (columns 1–3) and 2, respectively. We observe a statistically significant preference for the proposed system in the MUSHRA scores across all evaluated metrics. The objective results align with these findings, suggesting that the proposed method surpasses other speaker adaptation techniques when using the same amount of data.

Table 1. Average objective performance for all evaluated systems. The top performance is highlighted in bold. TTSuzb was trained on 25 minutes of speech rather than 1.

System				
	VF	*CC*	*MS-TTS*	*TTSuzb*
CSED	***0.191***	*0.198*	*0.205*	*0.207*
cFSD	***0.198***	*0.251*	*0.261*	*0.228*

Table 2. Average MUSHRA results with a 96% confidence interval. The top results, showing a statistically significant difference between the voice filter (VF) and reference systems, are highlighted in bold (p<0.1).

System					
	VF	*CC*	*MS-TTS*	*Rec*	*Lower-anchor*
Sp.sim.	***67.98***	*66.55*	*65.85*	*79.48*	*38.00*
Style sim	***64.6***	*63.56*	*62.13*	*75.57*	*38.05*
Nat.	***54.1***	*52.1*	*50.75*	*79.01*	-
Sig. Q.	***65.3***	*55.30*	*54.27*	*76.81*	-

IV.II. Ablation Study on Data Quantity

To understand the influence of the extremely low-resource scenario, we trained the Voice Filter using 1, 10, and 30 minutes of target speaker data during fine-tuning. The objective and perceptual MUSHRA scores are presented in Tables 3 and 4, respectively. In the subjective evaluation, listeners did not detect a statistically significant difference between the different data scenarios, although the objective scores still show slight improvements with larger data amounts. This indicates that while there is potential for enhancement in system performance, the Voice Filter does not perceptually benefit from richer data scenarios. With only 1 minute of target data, we are able to produce high-quality samples.

Table 3. Average objective performance for Voice Filter trained on varying amounts of data. The top results are highlighted in bold.

# min	***1***	***10***	***30***
CSED	*0.191*	***0.181***	*0.186*
cFSD	*0.198*	*0.189*	***0.175***

Table 4. Average MUSHRA results with a 96% confidence interval for Voice Filter trained on different data sizes. There are no statistically significant differences (p<0.1) between the systems.

# min	***Rec***	***1***	***10***	***30***
Sp.sim.	*74.01*	*51.93*	*51.94*	*52.07*
Style sim	*71.98*	*54.53*	*55.41*	*54.81*
Nat.	*79.00*	*54.99*	*55.11*	*55.19*
Sig. Q.	*74.02*	*54.03*	*53.93*	*53.65*

IV.III. Comparison with a competitive TTS

Finally, the quality of the generated speech is assessed in comparison to a TTS system trained on a larger set of target recordings. For this comparison, we evaluate the Voice Filter against a proven low-resource TTS technology that has demonstrated competitiveness when trained on 30 minutes of target speech (TTSuzb) [6]. It is important to note that such a technology implicitly leads the TTS system to estimate the duration of a phone call for the target speaker, which is not the case with the proposed Voice Filter. Essentially, this comparison pits the Voice Filter against a system trained on 30 times more target data, which has also shown competitiveness with TTS voices trained on 5+ hours of target recordings.

The objective metrics and MUSHRA perceptual scores are presented in Tables 1 (columns 1 and 4) and 5, respectively. Although we observe a relative degradation of 4% in speaker similarity, the MUSHRA scores do not indicate statistical differences between the systems in terms of signal quality, naturalness, and style similarity. On the other hand, the objective metrics reveal that speaker similarity and signal quality are superior with our proposed method. Overall, the results demonstrate that our model is on par with TTSuz, with only a slight human-perceived degradation in speaker similarity, despite the much smaller target training dataset used.

Table 5. Average MUSHRA results. The highest scores, showing statistically significant differences between the Voice Filter (VF) and reference systems, are highlighted in bold (p < 0.1).

System				
	VF	*TTSuzb*	*Rec*	*Lower-anchor*
Sp.sim.	*70.40*	***73.66***	*83.21*	*37.75*
Style sim	*69.55*	*70.01*	*79.81*	*39.54*

Nat.	*55.31*	*55.23*	*77.09*	-
Sig. Q.	*55.59*	*55.85*	*76.77*	-

In this study, we introduced a new, highly low-resource TTS method called Voice Filter, capable of producing high-quality speech using only 1 minute of audio data. Voice Filter divides the TTS process into two tasks: generating speech content and identifying the speaker. The speaker identification is achieved using a fine-tuned one-to-many VC module, which makes it easily adaptable to new speakers, even in settings with minimal resources. The speech content generation is handled by a duration-controlled single-speaker TTS system, which also facilitates the creation of a synthetic parallel corpus. This approach enables Voice Filter to function in a frame-level parallel environment, offering a higher potential quality and reduced modeling complexity.

Evaluations indicate that our Voice Filter surpasses other few-frame speech synthesis techniques in both objective and subjective metrics within the 1-minute data scenario, achieving quality levels comparable to state-of-the-art systems trained on 30 times more data. In conclusion, we consider the Voice Filter model a foundational step towards developing extremely low-resource TTS as a VC plug-in for post-processing. Additionally, we believe that the ability to generate synthetic parallel data with controlled duration will open up new possibilities in speech technologies that were previously constrained by data limitations.

CONCLUSION

In this study, we introduced a new, highly low-resource TTS method called Voice Filter, capable of producing high-quality speech using only 1 minute of audio data. Voice Filter divides the TTS process into two tasks: generating speech content and identifying the speaker. The speaker identification is achieved using a fine-tuned one-to-many VC module, which makes it easily adaptable to new speakers, even in settings with minimal resources. The speech content generation is handled by a duration-controlled single-speaker TTS system, which also facilitates the creation of a synthetic parallel corpus. This approach enables Voice Filter to function in a frame-level parallel environment, offering a higher potential quality and reduced modeling complexity.

Evaluations indicate that our Voice Filter surpasses other few-frame speech synthesis techniques in both objective and subjective metrics within the 1-minute data scenario, achieving quality levels comparable to SOTA systems trained on 30 times more data. In conclusion, we consider the Voice Filter model a foundational step towards developing extremely low-resource TTS as a VC plug-in for post-processing. Additionally, we believe that the ability to generate synthetic parallel data with controlled duration will open up new possibilities in speech technologies that were previously constrained by data limitations.

REFERENCES

[1] Kakhorov A.A., Yodgorova D.M., Khujakulov T.A., Bozorova Z.S. "Processing models and algorithms of natural languages"- Descendants of Muhammad al-Khwarizmi, №3(21)/2022.

[2] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, et al., "Effect of data reduction on sequence-to-sequence neural tts," in Proc ICASSP. IEEE, 2019, pp. 7075–7079.

[3] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in Proc. ICASSP). IEEE, 2019, pp. 6940–6944.

[4] Y.-J. Chen, T. Tu, C. chieh Yeh, and H.-Y. Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in Proc. Interspeech 2019, 2019, pp. 2075–2079.

[5] Q. Xie, X. Tian, G. Liu, et al., "The multi-speaker multi-style voice cloning challenge 2021," in Proc. ICASSP, 2021, pp. 8613–8617.

[6] Y. Chen, Y. Assael, B. Shillingford, et al., "Sample efficient adaptive text-to-speech," in International Conference on Learning Representations, 2019.

[7] Kakhorov, A. (2023). Og'zaki muloqot tizimlarini ishlab chiqish uchun noravshan qoidalarga asoslangan evolyutsion klassifikatorlarning qo'llanilishi. DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE, 1(2), 108–115. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v1i228

[8] D.-Y. Wu, Y.-H. Chen, and H. yi Lee, "VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net Architecture," in Proc. Interspeech 2020, 2020, pp. 4691–4695.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[10] S. Choi, S. Han, D. Kim, and S. Ha, "Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding," in Proc. Interspeech 2020, 2020, pp. 2007–2011.

[11] Y. Chen, Y. Assael, B. Shillingford, et al., "Sample efficient adaptive text-to-speech," in International Conference on Learning Representations, 2019.

[12] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in Advances in Neural Information Processing Systems, 2018, vol. 31.

[13] Kakhorov A., Yodgorova D., and Khayrullayev M.. "AI-driven environmental communication: translating sustainability reports into uzbek for global awareness" Экономика и социум, no. 6-1 (121), 2024, pp. 292-301.

[14] M. Chen, X. Tan, B. Li, et al., "Adaspeech: Adaptive text to speech for custom voice," International Conference on Learning Representations (ICLR), 2021.

[15] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, "Adadurian: Few-shot adaptation for neural text-to-speech with durian," arXiv preprint arXiv:2005.05642, 2020.

[16] H. B. Moss, V. Aggarwal, N. Prateek, J. Gonz´alez, and R. Barra-Chicote, "Boffin tts: Few-shot speaker adaptation by bayesian optimization," in Proc. ICASSP. IEEE, 2020, pp. 7639–7643.

[17] A.A.Kakhorov, D.M.Yodgorova, T.A.Khujakulov, Z.S.Bozorova Methods of using natural language processing for digital profile, Bulletin of TUIT: Management and Communication Technologies 2023.Vol-2(8)

[18] T. Wang, J. Tao, R. Fu, et al., "Bi-level speaker supervision for one-shot speech synthesis.," in Proc. Interspeech, 2020, pp. 3989–3993.

[19] Z. Cai, C. Zhang, and M. Li, "From Speaker Verification to Multispeaker Speech Synthesis, Deep Transfer with Feedback Constraint," in Proc. Interspeech 2020, 2020, pp. 3974–3978.

[20] Y. Jia, Y. Zhang, R. J. Weiss, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Advances in Neural Information Processing Systems 31, 2018, pp. 4485–4495.

[21] H. B. Moss, V. Aggarwal, N. Prateek, J. Gonz´alez, and R. Barra-Chicote, "Boffin tts: Few-shot speaker adaptation by bayesian optimization," in Proc. ICASSP. IEEE, 2020, pp.7639–7643.

[22] M. Ott, S. Edunov, A. Baevski, et al., "fairseq: A fast, extensible toolkit for sequence modeling," in Proceedings of NAACL-HLT 2019: Demonstrations, 2019

[23] M. Bi´nkowski, J. Donahue, S. Dieleman, et al., "High fidelity speech synthesis with adversarial networks," in International Conference on Learning Representations, 2020.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 12449–12460.

[25] Muminov, B., Nasimov, R., Gadoyboyeva, N. and Mirzahalilov, S., 2019. Estimation affects of formats and resizing process to the accuracy of convolutional neural network. In International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019 (pp. 9011858-9011858).