Available online at www.jpit.az

16 (2)
2025

# SARA: Synthetic Azerbaijani Recording Alignment – A Unified TTS Framework Trained on Synthetic and Real Speech

Mirakram Aghalarov[a], Kavsar Huseynova [b], Elvin Mammadov[c], Gurban Guliyev [d]

[a,b,c,d] Baku Higher Oil School, Yeni Salyan Highway 3rd km, 25, AZ1023 Baku, Azerbaijan

[a] mirakram.agalarov@bhos.edu.az; [b]kavsar.huseynova.std@bhos.edu.az; [c] elvin.mammadov.std@bhos.edu.az; [d]gurban.guliyev.std@bhos.edu.az

[a] https://orcid.org/0009-0008-1551-7797; [b] https://orcid.org/0009-0007-0362-9591; [c] https://orcid.org/0009-0005-9237-9736 [d]https://orcid.org/0009-0004-6368-2720

ARTICLE INFO

**Keywords:**

Speech Synthetis
Azerbaijani TTS
Deep learning
Voice Processing
Artificial Intelligence

ABSTRACT

Speech synthesis is crucial part of the modern assistive AI (Artificial Intelligence). While the development in this field has great burst, solutions for low-resource languages stays in background due to lack of the data. Our paper demonstrates the novel solution to overcome this problem for Azerbaijani Language and proposes a model which can generate fluent voice in Azerbaijani Language. Our solution is based on self-supervised learning with multi-lingual Text-to-Speech (TTS) model to pretrain the VITS (Variational Inference for Text to Speech) architecture from scratch. Additional Fine-tuning dataset has been obtained by segmenting voiced news articles along with text of the news by using whisper. Moreover, obtained datasets for fine-tuning and pre-training has been made available for further development and use-cases as an open-source. Model has been evaluated according to the LLM (Large Language Models) and Human Opinion by using Mean Opinion Scores (MOS) which proved efficiency of the approach and usability of the proposed model for further applications. Model is available at: https://huggingface.co/BHOSAI/SARA_TTS

## 1. Introduction

Voice processing is part of the Deep learning which is studied for long time to create applications like Automatic Speech recognition (ASR), speaker identification, noise removal and Speech synthesis (Khanam et al. 2022). All these applications were widely used in production such as call centers, synchronous translation. Maturity of the applications depend on maturity of datasets which the models were trained with.

Selection of the dataset is paramount task for the Text to Speech Synthesis (TTS) as it should sound human. Additional challenge is that TTS applications are language specific which means that models are not suitable to the multilingual setup. This creates a bottleneck when new model should be built for the language which is considered as low-resource. These languages do not have enough dataset to train the model and requires annotation based on factual data.

Azerbaijani language is considered as low-resource language. It is very difficult to find audio-label pairs for clean data containing low number of speakers. Considering the cost of the annotation or voice recordings, it is more challenging to find the data in open-source.

There are some models in open-source which are published by Meta (Vineel Pratap et al. 2023) to be used in local machines. As they don't have any labelled dataset, methods like zero-shot learning were very successful to build those

*Received 26 February 2025, Received in revised form 29 April 2025, Accepted 13 May 2025*
http://doi.org/10.25045/jpit.v16.i2.07

models with large amount of unlabeled data. Considering that these methods require high computational resources, replicating this experiment on more accurate voices is almost impossible while replication of experiment by natives would increase the quality by much.

We are introducing novel pipeline SARA (Synthetic Azerbaijani Recording Alignment) and open-source model with dataset to overcome this challenge for low-resource languages focusing on Azerbaijani. We developed a pipeline to utilize self-supervised learning on pre-training from already trained model from Meta. This step allowed us to give the model with large amount of the data to teach base phonetics of Azerbaijani language. Moreover, we have utilized and processed the audio news from the news agency website to create sentence-based voices with corresponding content.

Creation of the dataset based on news agency helped us to diminish the time to be spent on manual annotation and increased the quality as in news sessions, the speech should be grammatically very accurate and the speakers are less varying.

Research has been done in several step:

- Collection of the data: 3 methods have been used – crowd-sourcing from university students, pseudo labels from trained TTS model, validation with content based on automatic speech recognition of the audio news.
- Selection of the model: Literature review has been done to select what type of the model should be used.
- Training with the novel pipeline: Self-supervised training on pretraining and news data in fine-tuning.

The paper is organized as follows. State-of-the-art TTS methods and models for especially low-resource languages are analyzed and compared in Section 2 – Related Work. Details of the knowledge distillation method which delivered robust model has been described in 3rd section while the results, comparison of the state-of-the-art and discussion can be observed in Section 4 – Experimental Results. Research has been concluded in Section 5 and development for the future has been discussed.

## 2. Related work

**Model architectures:** TTS models gave stunning results in last 10 years. Tacotron (Yuxuan Wang et al. 2017) was most significant research in this direction. Model was developed to convert the given text in natural language into spectrogram. Attention mechanism inside the model was passing the features extracted into autoregressive decoder RNN (Recurrent Neural Networks) which generated mel-spectrogram. Considering that the generation of the audio out of mel-spectrogram is not linear process (audio has more features than spectrogram) vocoder is used to generate the audio.

After autoregressive model, FastSpeech (Ren et al., 2019) developed non-autoregressive model based on feed-forward transformer encoder to generate hidden states from phonemes of the language. Main difference in this model is that the model was able to generate frames in parallel which increased the speed of the process.

Glow-TTS (Kim et al., 2020), has latent features aligned with generated spectrograms. As FastSpeech, it is able to generate frames of the spectrogram in parallel which gives more stable generation while strictly advised that training must be monitored carefully. Computational cost of the training is very high due to the model complexity and inference poses trade-off between computation and quality.

VITS (Kim et al. 2021), proposed unified approach to generate directly the waveforms. Methodology includes adversarial learning by applying discriminator to the output of the model which can diversify between human generated and model generated responses. It outperforms all 2 stage TTS architecture while increasing the training complexity.

**TTS models for Azerbaijani**: Considering that Azerbaijani language has limited digital language corpora, development in natural language processing application are very slow. (Valizada et al. 2x022) has collected 24 hours of recording from single speaker and trained their Tacotron2 model on it. However, the dataset was not found publicly on open-source therefore experiment could not be replicated.

(Zeynalov et al. 2024) collected 11 hours of recording from internet and less than 1 hour of recording from call center recordings to train their Tacotron 2 model with Vocoder based on HifiGAN. Authors have achieved MOS = 4.17.

**TTS training for Low resource languages:** (Kim et al. 2022) proposed training VITS model with 1000+ hours of unlabelled data using self-supervised techniques to derive "pseudo-phonemes". Then they used only 10 mins of the

clean audio to fine-tune.

Very same paper (Kim et al. 2022) proposes another approach based on zero-shot learning to mimic the sound of the person given as an input. They propose that few minutes of language specific audio in multilingual model could bring better performance to the model in low-resource setting.

## 3. Proposed approach

Considering that Azerbaijani language is low resource language, dataset building becomes more crucial step for successful training. During research, the first dataset found for the training was from Mozilla commonvoice where the data is in pairs with sound and corresponding text. Although the dataset is designed for Automatic Speech Recognition collected by crowd sourcing, number of speakers in given dataset is very low. However, size and quality of the dataset is not sufficient for pretraining.

In order to overcome this issue, first we selected training methodology in low-resource environment. Self – Supervised Learning for pretraining is the best choice when there is unlabelled data and teacher model. After Self-supervised learning in order to increase the naturalness of the model there is need for fine-tuning with very clean and less amount of data.

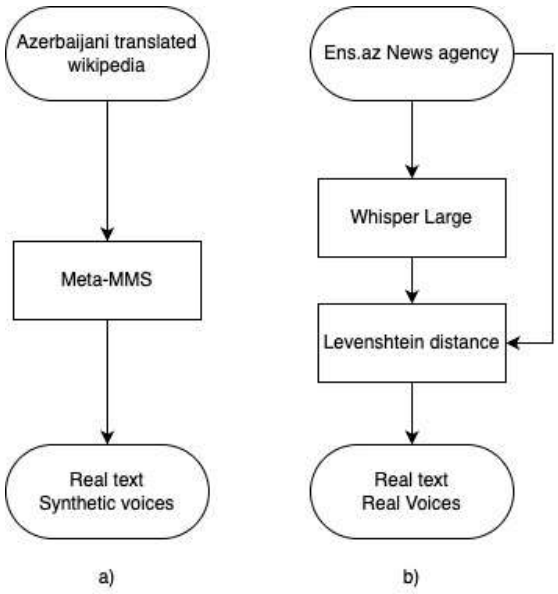


**Fig. 1.** Creation of the dataset: a) Pre-training dataset from pseudo-labels. b) Fine tuning dataset processed by Whisper Large

As a result of planning, 2 datasets have been created:

1. Pretraining dataset – Collected using MMS (Massively Multi-lingual Speech) model by Meta (Vineel Pratap et al. 2023) with Azerbaijani version. We used data from Azerbaijani wikipedia dataset from Wikimedia, as text source, to generate the voices from TTS model. Despite the fact that MMS model by Meta does not have stable performance on Azerbaijani, high amount of text inputs would create general representation of the Azerbaijani TTS. 160k voices was enough to perform knowledge distillation which would give good pretraining to be able to analyze and process phonems and letters.
2. Fine-tuning dataset – this dataset is created for fine tuning after self-supervised learning. Dataset has been collected from ens.az news agency which also supports audio news with corresponding transcriptions in website. Considering that news agencies have very clean speech and less number of various speakers, utilization of this data for training would boost up the performance. However, problem arose when the data size was longer that 30 seconds which was frequent as they are news explanations. We used whisper model by OpenAI (large) in order to transcribe all audios from new agency into text. Then, we used levenshtein distance in order to compare the original transcriptions with the prediction as Whisper model can have some errors. Most compatible sentence from the news is selected as ground truth of the given audio. By this way, we created 3.5k voice-text pairs.

According to the Literature Review we have selected VITS model for training considering naturalness criteria which comes from GAN (Generative Adversarial Networks) inside. Unlike other TTS models, VITS has only one processing unit during inference which decreases the computation comlexity.

In the first step of the training, we have used pre-training dataset to acknowledge VITS model with the structure of Azerbaijani Language. The pre-training has been done using knowledge distillation technique which helped to generalize problems that MMS models experienced. Therefore, the newly pretrained model created was still better than the MMS model.

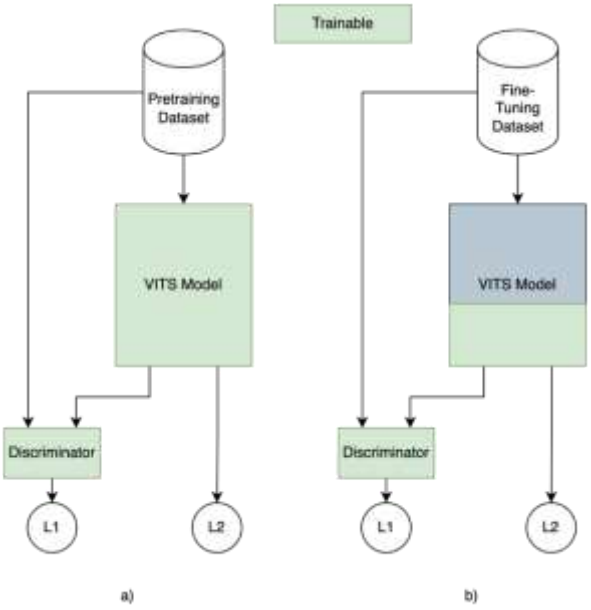


**Fig. 2.** a) Pretraining with full parameters activated. b) Fine-tuning with half parameters activated

Secondly, we used the 2 types of fine-tuning dataset in which we failed with one of them. We used crowd-sourcing approach from university with telegram-bot to collect data. However, quality of the recording and consequently the training did not give desired result. On the contrary, dataset based on news agency gave perfect results which is discussed in experimental results section.

## 4. Experimental results

All experiments have been conducted in multi-node single-GPU setup. 5 nodes were equipped with RTX4090 with high performance network interface card to avoid network bottleneck.

As mentioned in methodology section, 2 datasets have been used for this operation. 1 for pretraining of the VITS, another for fine-tuning the pretrained model in Azerbaijani.

Table 1 describes the details of the pretraining and finetuning dataset. Generally multi-speaker datasets are used for pretraining and single speaker for the fine-tuning. However, in our case, we had to take into account that synthetic dataset can't be used for the fine-tuning and large number of samples are needed to be for pretraining. Therefore, Synthetic dataset with 162k samples by single speaker (model) is used as pretraining and real dataset with 3510 samples by multiple speakers is for the finetuning.

**Table 1.** Description of the used datasets

| Dataset | Number of samples | Multi-Speaker |
|---|---|---|
| Synthetic Dataset from MMS model | 161,955 | no |
| Fine-tuning dataset from ens.az news agency | 3510 | yes |

According to the experiments, usage of the dataset has been changed several times. When single sentences have been used, sounding of the sentence did not feel like sentence ending. Therefore, we decided to extend the voices to 2 and 3 voices. According to the experiments, we also utilized random allocation of the number of sentences in given input data as in described Table 2.

**Table 2.** Parameters used during experiments

| Parameter | Value |
|---|---|
| Number of Sentences in 1 audio | 1 / 2 / 3 / 1,2,3 |
| Number of epochs | 120 |

To understand the experimentation results, we used several metrics like MCD score (Mel-Ceptral Distortion), SNR (Signal to Noise Ratio) and MOS (Mean Opinion Score). MCD gives us notion about alignment of the spectrum, while Signal to Noise Ratio indicates how data is distributed with noisy frequencies. MOS is based on Human opinions about 2 characteristics: Intelligence and naturalness.

Mel-Ceptral Distortion (MCD) is given as:

$$MCD[dB] = \frac{10}{\ln(10)} * \sqrt{2} * \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{d=1}^{D} \left(c_{t,d} - C_{t,d}\right)^2}$$

Where:

- T is the total number of frames,
- D is the number of MFCC coefficients (excluding the 0th coefficient in many cases),
- The constant $\frac{10}{\ln(10)} * \sqrt{2} \approx 6.14185$ converts the Euclidean distance to decibels (dB).
- $c_t = [c_{t,1}, c_{t,2}, c_{t,3}, c_{t,4}, c_{t,5}, \ldots. c_{t,D}]$ reference (e.g., ground truth) MFCCs at frame t
- $C_t = [C_{t,1}, C_{t,2}, C_{t,3}, C_{t,4}, C_{t,5}, \ldots. C_{t,D}]$ synthesized or predicted MFCCs at frame t

Signal to Noise ratio is calculated in decibels (dB) for our experiments as in given below:

$$SNR_{dB} = 10\log_{10}\left(\frac{\sum_{n=1}^{N} x[n]^2}{\sum_{n=1}^{N}(x[n]^2 - \hat{x}[n]^2)}\right)$$

Where:

- $x$[n] – Ground truth signals (reference audio)
- $\hat{x}$[n] – Generated signals (prediction audio)
- N – Total Number of samples in signal

By Signal to Noise ratio difference between ground truth and prediction can easily be found as formulation explains division of the original audio by error between original and generated samples. In the context of TTS, a higher SNR indicates that the synthesized waveform closely matches the reference, implying less distortion or noise introduced by the synthesis model.

Mean Opinion Score (MOS) blind evaluation of the audio by certain group of people. Audios are given in random manners and evaluators are asked to score them according to 2 criterias. Intelligence and Naturalness. Both criterias are scored between 1 and 5.

- Intelligence – Correct sounding of phonemes, start and end of the sentences, commas.
- Naturalness – If sound of the speaker seems realistic or not.

**Table 3.** Results of the pre-training. After pretraining to assess the quality, we have tried same set of evaluations

| Model Name | MCD | SNR | MOS | |
|---|---|---|---|---|
| | | | Intelli. | Natural. |
| Tacotron 2 | 3.9 | 32 | 2.3 | 2.8 |
| VITS | 4.0 | 33 | 2.1 | 2.9 |

**Table 4.** Results of the fine-tuning after several combination of experimentation

| Model Name | Num of Sentences | MCD | SNR | MOS | |
|---|---|---|---|---|---|
| | | | | Intelli. | Natural. |
| Tacotron 2 | 1 | 6 | 29 | 3 | 3.6 |
| Tacotron 2 | 2 | 4.2 | 26 | 3.1 | 3.9 |
| Tacotron 2 | Random | 4.4 | 30 | 3.8 | 4.1 |
| VITS | 1 | 4.3 | 30 | 3.5 | 4.0 |
| VITS | 2 | 4.3 | 31 | 3.3 | 4.0 |
| VITS | Random | **4.1** | **31** | **4.1** | **4.3** |

According to the Table 3 it is clearly seen that Tacotron Model is not performing well for our data despite higher results in global benchmarks. The reason is that, Tacotron 2 is susceptible to the noises in our dataset as it is scraped from news website in which the data has 128 kbs. This is crucial factor as it shows its effect in the output.

Table 3 proves the quality difference between pretraining and fine-tuning datasets as synthetic dataset had much cleaner outcomes. Therefore, we see better results for the Tacotron 2 model in MCD and SNR scores. Human evaluators did not find naturalness of the outcome well enough. This is because the teacher model (MMS by Meta) was not natural enough. The main purpose of the pretraining was to establish robust feature extractor with tokenizer which could influence to the fine-tuning results.

Another crucial outcome from Table 4 is that single sentences are not enough to express the intonation. Therefore, Random selection of number of sentences increased quality of the audio. Both models get benefit from the random selection and as score is increasing. VITS show higher SNR than the Tacotron 2 model while also combination of the sentence had achieved cleaner outcomes with better spectral Alignment (MCD). Human Reviewers felt that most natural and accurate model was VITS with random multiple sentences with special note that, “models trained with singular sentences did not generate natural ending for the sentence and felt like it will go on”.

## 5. Conclusion

This paper represented the work done around Text to Speech Synthesis for Azerbaijani Language. Comprehensive approach with pretraining and additional fine-tuning showed its effect on the end results. Human evaluators evaluated with blind review without knowing that if the audio is coming from TTS model or if it is ground truth. Results show effectiveness of the methodology, usability in real-world cases. Open-source publish of the model encourages the researchers to take the model and the dataset for further development.

Our methodology is proved to be effective in language agnostic settings. The same pipeline can be used to train the TTS models in other low resource languages without change. Considering the 2-step training pipeline, resource allocation was high considering that the pretraining was

costly and utilized 3 out of 5 RTX4090 to implement good generation capability.

As for future development, model should be extensively tested in real cases to understand the weaknesses in generation. Some abbreviation in English languages can be included in pretraining datasets to increase use cases for the TTS. Additionally, fine-tuning dataset can be more cleaned considering that original kbps (128) was not considered high quality enough for smoother generation. From methodology side, mini-LLM based approach could also increase the naturalness with emotion prediction.

## References

Khanam, F., Munmun, F. A., Ritu, N. A., Saha, A. K., & Firoz, M. (2022). Text to speech synthesis: A systematic review, deep learning based architecture and future research direction. Journal of Advances in Information Technology, 13(5), 398–412. https://doi.org/10.12720/jait.13.5.398-412

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the International Conference on Machine Learning (ICML) (pp. 5530-5540). https://doi.org/10.48550/arXiv.2106.06103

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. Proceedings of the 34th İnternational Conference on Neural Information Processing Systems (pp. 8067–8077). https://doi.org/10.48550/arXiv.2005.11129

Kim, M., Jeong, M., Choi, B. J., Ahn, S., Lee, J. Y., & Kim, N. S. (2022). Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus. Proceedings of Interspeech 2022 (pp. 788–792). https://doi.org/10.21437/Interspeech.2022-225

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25, 1-52. https://doi.org/10.48550/arXiv.2305.13516

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). FastSpeech: Fast, robust and controllable text to speech. Proceedings of the 33rd International Conference on Neural Information Processing Systems (pp. 3171-3180). https://doi.org/10.48550/arXiv.1905.09263

Valizada, A., Jafarova, S., Sultanov, E., & Rustamov, S. (2021). Development and Evaluation of Speech Synthesis System based on Deep Learning Models. Symmetry 2021, 13(5), 819. https://doi.org/10.3390/sym13050819

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017) Tacotron: Towards end-to-end speech synthesis. Proceedings of Interspeech 2017 (pp. 4006-4010). https://doi.org/10.21437/Interspeech.2017-1452

Zeinalov, T., Sen, B., & Aslanova, F. (2024). Text-to-Speech in Azerbaijani language via transfer learning in a low resource environment. Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 434-438). https://aclanthology.org/2024.icnlsp-1.44/