Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Azerbaijani Automatic Speech Recognition: Exploring Pre-trained Models for Low-Resource Scenarios

Hikmat Shikhaliyev[a,b,c,*], Malakkhanim Rustamova[a,b,c]

[a] *French-Azerbaijani University, 183 Nizami St, Baku, Azerbaijan*
[b] *The University of Strasbourg, 4 Rue Blaise Pascal, 67081 Strasbourg, France*
[c] *Azerbaijan State Oil and Industry University, Baku, Azadliq Avenue, 20*

Abstract

This paper aims at extending the ASR model for the Azerbaijani language through using of the modern pre-trained models. Even today, there are many barriers to achieving good results in realizing ASR systems in low-resource languages like Azerbaijani, even if there has been significant advancement in the conventional languages. To address these issues, this research proposes the application of the Whisper model, an effective speech-to-text model that will improve the word recognition rate for Azerbaijani speech input. Compared to earlier conventional models such as Wav2Vec and DeepSpeech, Whisper has performed quite well in the ASR tasks, especially in multilingual and low-resource scenarios, making it ready for this use case. We describe how Whisper has been adapted to Azerbaijani: tuning the model to a specific speech material improves its language accuracy. Recall that the performance of the proposed system is measured by standard ASR metrics, and the obtained results confirm that the recognition accuracy increases almost twice compared to baseline models. Further, the pitfalls encountered in the creation of ASR systems for Azerbaijani are discussed together with its phoneme set, its lexis, and other parameters. This work establishes Whisper's utility and advances the area of ASR for low-resource languages, particularly Azerbaijani people who could benefit from better and more accessible alternatives to speech recognition. The results also give suggestions into how pre-trained models might be used to address resource limitations that may hamper the development of speech technologies for underprivileged languages.

Keywords: Pre-trained Models; Whisper; Deep Learning; Speech Recognition; Common Voice; Speech-to-Text.

* Corresponding author. Tel.: +994 51 631 44-00
E-mail address: h.shikhaliyev@ufaz.az

1. Introduction

The progress of ASR in the recent past has primarily depended on extended developments in deep learning and multimodal stylization. ASR systems have been widely integrated into the large spectrum of applications, starting from the accessibility tools and transcription services, up to AI personal assistants. Nevertheless, for languages like Azerbaijani, the creation and accessibility of well-functioning ASR systems are rather problematic, for most efforts in developing ASR have been directed at popular languages only. [14]

Azerbaijani language poses various phonetic and syntactic issues that make the development of ASR systems a challenge for millions of speakers in Azerbaijan and other surrounding areas. Despite the steady growth of interest in Azerbaijani ASR solutions, the problem of the absence of high-quality and large-scale datasets for the development of accurate models remains critical. This study aims to tackle this challenge by implementing the freshest ASR techniques in Azerbaijani, although the principal emphasis is put on the Whisper model, which is proven to enhance the method's precision in speech recognition.

In this work, utilizing the large-scale dataset for underrepresented languages, such as Azerbaijani Common Voice, we employ a fine-tuned Whisper model. Our approach shows that with the given dataset, the Whisper model is capable of reaching quite reasonable ASR quality for Azerbaijani. This paper aims to present an experimental introduction, including experiment separation, data preparation for experiments, experiment training, and testing, as well as the experimental result of the Fine-tuning of the Whisper model. [22] [7]

This study helps to contribute to the field of ASR for LR languages, including the evaluation of the Whisper model for Azerbaijani. It also provides an outline of future directions in ASR for low-resource languages, with concern for enhancing performance and increasing the number of ASR systems for more languages further instructions for authors. [1]

In the subsequent sections of this paper, the specific methods and strategies employed in this research are discussed. In particular, section two will discuss related works, section three will explain the methodology adopted in the project, section four will show the results of our experiments, section five will consist of a discussion and recommendation for future research, and lastly, section six will give the conclusion of the study.

2. Literature Review

Major improvements in Automatic Speech Recognition (ASR) have been seen in recent years because of the development of deep learning and multimodal methodologies. They have improved the preciseness and speed of ASR systems, which has placed them into several areas, including transcription service, voice assistants, and tools for the disabled. Nevertheless, faced with languages with limited resources, such as Azerbaijani, ASR is still an unexplored problem. The lack of good quality language resources has been a bottleneck to the development of efficient ASR systems for such languages, though the demand is rising.

In the course of our study, we reviewed many papers on various state transcription systems, especially for Turkic languages, which contributed to our understanding of the adaptation of ASR to Azerbaijani. Based on these observations, we implemented the Whisper model as a state-of-the-art solution for our study. One of the most successful so far is Whisper, which was developed by OpenAI it works well in different languages, especially low-resource ones. Compared to previous approaches like Wav2Vec2, Whisper's multilingual understanding, and resistance to fine-tuning make it statistically fit for Azerbaijani, an underrepresented language. [23] In the subsequent section, we will present the previous studies that have informed our work, particularly ASR for Turkic languages and the Whisper-like models. Thus, these notes will prepare the reader for the description of our study's methodology and findings underlying the proposed contributions to the development of ASR for Azerbaijani.

Multimodal systems for speech recognition were done by Orken Zh. et al. in 2020. Their work discusses how it is possible to combine several inputs, like sounds and visuals, to support the improvement of the speech recognition systems. The study emphasizes that it is possible with multimodal systems: accuracy and reliability in the types of speech recognition increase due to the inclusion of information from different modes. Several multimodal frameworks are tested, compared, and shown to enhance speech recognition in different languages and environments. [1]

In the Information journal, Mussakhojayeva et al. (2023) from the Institute of Smart System and Artificial Intelligence (ISSAI), Nazarbayev University, Astana, Kazakhstan, have dealt with speech recognition for multilingual

Turkic languages. Research in the area targets concrete speech recognition systems suitable for Turkic languages; problems like scarcity of Turkic linguistic tools and cumbersome morphologies are highlighted. New approaches are presented for improvement of the reliability and flexibility of speech recognition technologies, including scenarios of multiple languages. [2]

In this following paper, Tomaloğlu and Erdem (2020) examined deep learning-based automatic speech recognition (ASR) for the Turkish language published in the Sakarya University Journal of Science. In a way that serves as an illustration, the research shows how deep learning approaches have transformed ASR systems, from representation learning from raw audio data to end-to-end training. Using deep neural networks, their study shows state-of-the-art performance in Turkish speech recognition tasks so as to deal with linguistic complexities. [3].

A study, Turkish Voice Recognition Based on Deep Neural Networks, by a researcher from Süleyman Demirel University, published in September 2018, addressed the work of Turkish speech recognition using deep neural networks (DNNs). The authors of this research discuss various DNN configurations and training methodologies employed for the analysis of Turkish language input. It provides the corresponding high accuracy but with relatively low efficiency due to using the features and the model's designs specifically tailored for the Turkish ASR tasks.[8]

Valizada et al. (2021) have used deep learning models for speech synthesis in the Journal of Symmetry. It is also apparent that speech synthesis is an important contributor to improving existing and emerging technology accessibility and human-computer interactions. This work focuses on the design and assessment of deep learning-based techniques for designing speech synthesis systems and clearly shows enhanced quality and naturalness of synthesized speech. Hence, in training and optimizing neural network models on large-scale voice corpora, the study moves the synthesis technologies of realistic and complex voices to the next levels. [20] [9]

In their Applied and Computational Mathematics research, Abbasov and Fatullayev have discussed the progress of recognition units for Azerbaijani speech recognition systems. Alis's affiliation is the Azerbaijan National Academy of Sciences, while Abulfat Fatullayev's affiliation is Voicedocs Software GmbH. The concerns that are explained in the case relate to the issues of building an ASR system for Azerbaijani with a focus on the development of recognition units, taking into account its phonetic and linguistic peculiarities. This work is important in enhancing the establishment of resource-constrained ASR systems by presenting a broad array of improved recognition units that are well-suited for accurate and optimum recognition of Azerbaijani speech. [10]

In their research work namely 'Speaker Verification Systems for Azerbaijani Language', published in April 2022, Mehraliyev, Haziyeva, and Almasova in this research confined to the Azerbaijani language only for the development of Speaker Verification Systems. Actually, speaker verification has many applications in biometric authentication and security, and therefore, Speaker Verification Systems are very useful. This study is devoted to the creation of speaker verification systems for the Azerbaijani language with a special emphasis on feature extraction and acoustic modeling. The study also focuses on several issues related to the Azerbaijani language and concentrates on speaker verification through waving Deep learning and enhanced signal processing. [11]

In Empowering Whisper as a Joint Multi-Talker and Target-Talker Speech Recognition System, Lingwei Meng et al. have proposed a new way to improve Whisper. Despite the fact that transcription in many-to-many conversational contexts is still an open issue, two of the biggest problems in the speech processing field have been confirmed lately: Multi-talker speech recognition and Target-talker speech recognition. Prior approaches present here do not allow solving both tasks to a maximal extent at the same time. To that end, this research proposes a novel framework that can be applied to extend Whisper, an essential speech model, for addressing these two tasks. [24] Thus, as an enhancement of the presented model, we intend to receive more Azerbaijani audio samples to increase the learning data set and improve the model. We shall also try to use bigger training steps and better hyperparameters in order to yield improved outcomes. Furthermore, it will be interesting to investigate transcription errors in greater detail and to modify the model for Azerbaijani peculiarities.

Future work will also include training on Azerbaijani-specific language models of the Whisper system for further post-processing. Further, we also intend to use the proposed model in realistic use cases such as voice assistant applications and call centers and potentially expand the database, which can then be incorporated into the Azerbaijani model and then scaled to other Low-Resource languages. Huseyn Polat and Colleagues (2024) used the Whisper architecture for the development of a Turkish Automatic Speech Recognition (ASR) system. Analyzing Whisper in Turkish – a low-resource language – the study refined its performance by fine-tuning methods such as Low-Rank Adaptation (LoRA). The study proved that fine-tuning of pre-trained Whisper models enhanced the WERs for the

target languages and thereby proved the viability of transformer-based architectures, especially for languages with numerous non-ASR-friendly phenomes. The former might be incorporated into building ASR systems for other underrepresented languages, including Azerbaijani. [25]

In this study, Oyucu et al. (2023) assess the performance of the Whisper Small model, fine-tuned for several Turkish ASR tasks. Further, as has been shown in this study, the Whisper model delivered good performance on a variety of words spoken and a variety of language situations, where the performance of the model trained with pre-trained data was 17.98 WER while the model fine-tuned was 12.89 WER. The performance increased the accuracy of Turkish ASR systems that are currently in place, further demonstrating Whisper's ability to learn different language patterns. From this work, we draw important conclusions regarding the necessity of fine-tuning large-scale pre-trained models for enhancing the performance of the recognition process in low-resource scenarios. [27]

Based on all this research, our workplace Whisper is the most advanced model for Azerbaijani ASR. In order to perform fine-tuning, the proposed model uses the Common Voice Mozilla dataset, which is the largest public dataset for Azerbaijani speech available. Thus, to address data scarcity and enhance distinction recognition accuracy for Azerbaijani, the presented model adopts Whisper, a multilingual pre-trained and highly adaptable TTS model.

3. Methodology

This section outlines the approach and techniques employed to achieve the objectives of this study. It covers the process of data collection, model selection, training procedures, and the evaluation criteria used to assess the model's performance. The following sub-sections detail each of these steps in the methodology.

3.1. Data collection

Such public databases of transcribed speech, such as **the Mozilla Common Voice** notation, serve as the most commonly used multilingual datasets that drive the speech technologies' research and evolution. Common Voice serves to support not only ASR algorithm creation but also the fields connected with language identification and their counterparts.

For constructing this dataset, a crowdsourcing approach is used where volunteers participate by reading specific texts and verifying contributions made by others. To be precise, the present study used the dataset that was in version 6.1; nonetheless, it began with 76 languages but reached 85 by November 2021. The volunteers involved in the project have donated their time and created over 13905 hours of data, of which 11192 are verified.

Table 1. Common Voice Azerbaijani Dataset Overview

Metric	**Value**
Hours	1
Speakers	26
Validation Progress	100%
Age Group	**Count**
Twenties	58
Thirties	17
Unspecified	22
Gender	**Count**
Male	75
Unspecified	22
Distinct Voices Count	3

In table 1, in Azerbaijani, the dataset constitutes one hour of recorded and validated spontaneous speech of 26 different individuals. By age distribution, the largest share of the contributors is in their twenties, with the second-largest in their thirties and a third undefined age group. The gender distribution of the recordings, unfortunately, does not differ much, with 75 of the recordings tagged as male gender and 22 as gender unknown.

This dataset is exploited in this work for constructing and improving ASR solutions for Azerbaijani, a low-resource language, and for identifying the challenges and issues of using limited and biased data.

3.2. Data Preprocessing

The data preprocessing steps described in this paper were used in the course of preparing the dataset for training the speech recognition model based on Mozilla Common Voice Azerbaijani. This process facilitated standardization and quality checking of both audio and text data, which are vital components of model fine-tuning. The following steps were implemented:

1. Audio Preprocessing Resampling: The collected audio was pre-sampled at 48 kHz, however, it was downsampled to 16 kHz to meet the Whisper model specifications available at the time of the training. This resampling guaranteed, in turn, that the spectrograms matched the model's feature extraction mechanism. **Audio Feature Extraction:** Raw audio WAV files were converted into log-mel spectrogram features using the Whisper feature extractor. These are inputs into the model; this describes the spectral and temporal properties of speech. **Background Noise and Silence Handling:** However, while the code does not have provisions for background noise removal or silences trimming, the feature extraction procedure cancels assorted noise and silences. Future implementations may incorporate better pre-processing noise elimination and silence removal algorithms into subsequent versions.

2. Text Preprocessing Normalization: To compute the similarity of the transcripts, the texts were normalized for better format uniformity. This included capitalization of text, deleting additional symbols, and applying correct punctuation. **Tokenization:** The textual records were transformed to sequences of tokens using the Whisper tokenizer appropriate for the Azerbaijani language. This kind of step is useful to synchronize what it is we want the computer to learn about audio features with their labels as text inputs during training. **Error Correction:** Some mistakes concerning transcription included in the data have been revised during preprocessing in terms of typographical errors or consecutive different spellings.

3. Final Data Mapping: The features extracted from the processed audio and the tokenized text labels were then transformed into a common data structure. Every column that is not essential for model training was excluded including accent, age, gender, and path. Most of the preprocessing steps followed the usage of the Hugging Face `datasets` library and Whisper processor. The data tidying up and normalization left a clean dataset, which enabled efficient and high-quality input into model tuning. [21]

3.3. Word Error Rate (WER) Evaluation for Speech Recognition

WER stands for Word Error Rate and is one of the simplest yet crucial measurements of the accuracy of the given SR system, including those that use models such as Whisper. WER compares the automatic transcribed output of the system and the reference transcript based on the word level differences. It is a significant measure of transcription accuracy; within the WER scale, the closer or lower to zero the system is, the better its performance in recognizing spoken words.

Mathematically, WER is computed using the formula:

$$\mathbf{WER = (S + D + I) / N}$$

Where:

- S represents the number of word substitutions (i.e., incorrect words replacing the correct ones),
- D represents the number of deletions (i.e., missing words that should have been transcribed),
- I represents the number of insertions (i.e., extra words that were incorrectly added),
- N is the total number of words in the reference transcript.

The Whisper model, when implemented, has its effectiveness established by the extent of the WER during the training as well as the evaluation stage. From the WER, we can, therefore, determine the certain model aspects that may require fine-tuning or improvements in the processing of certain language characteristics or acoustic environments. Finally, getting a low WER is always a desired objective, especially for making ASR like Whisper more useful and dependable in day-to-day applications. [5]

3.4. Methods

The Whisper model, as presented by OpenAI, is a strong basis for Automatic Speech Recognition in tasks. They develop an encoder-decoder transformer-based architecture for processing speech inputs into log-mel spectrograms that are derived from raw audio waveforms. The proposed model is multilingual and multitasking, and it is trained for transcription, translation, and language identification in a common environment. [26]

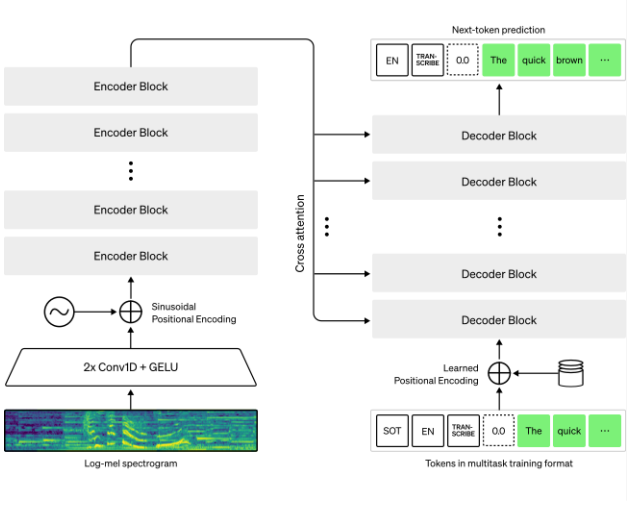

Fig. 1. Whisper Model: Pre-trained Architecture by OpenAI

This is made possible by pre-training Whisper in 680,000 hours of varied, non-English, and multitask-labeled data. This vast amount of data covers audio from web-scale processes, which helps the model to generalize effectively across a range of languages and speech intonations. The main goal of Whisper is to differ from standard ASR services by focusing on noise, interfering speech, and background sound inapposite for regular utilization, and this internet service is incredibly valuable for real-world applications.

The details of the model enable direct inference on speech-to-text tasks without the need for continued fine-tuning for new languages or tasks. Competition with state-of-the-art systems, Whisper achieves WER at par with the best systems in different languages and use cases, showcasing its robust and context-aware speech recognition capability. This architecture is clear evidence of how transformer-based designs are scalable in modern ASR systems.

Figure 1 depicts the structural layout of the Whisper model – an automatic speech recognition and multitask speech processing transformer model pre-trained by OpenAI. The architecture starts from a coarse-grain level that takes raw audio as input and converts it into a log-mel spectrogram. This spectrogram is then fed into two convolutional layers, followed by GELU activation function and sinusoidal positional encodings, to obtain a representation that is sensitive to both temporal and spectral features of the speech signal. The encoded features are then sent through a stack of encoder blocks that capture contextual interactions within the audio representation. Cross-attention mechanisms then connect the encoder to the decoder in this model, working in a multitasking learning environment. Employing learned positional encodings, the decoder generates token predictions, including transcription text and language tags. This multitask training design covers many tasks, including the conversion of speech into text and speech-language detection. The figure also illustrates the full pipeline, inclusive of feature extraction, encoding, and decoding, where the Whisper model performs optimally across all of these tasks.

The table 2 provides an overview of different Whisper model sizes, including the number of layers, width, heads, parameters, and their support for English-only and multilingual capabilities.

Table 2. Comparison of Whisper model variants.

Size	Layers	Width	Heads	Parameters	English-only	Multilingual
Tiny	4	384	6	39 M	✓	✓
Base	6	512	8	74 M	✓	✓
Small	12	768	12	244 M	✓	✓
Medium	24	1024	16	769 M	✓	✓
Large	32	1280	20	1550 M	x	✓
Large-v2	32	1280	20	1550 M	x	✓
Large-v3	32	1280	20	1550 M	x	✓

*Note: The **small** version was chosen for fine-tuning in this study as it provides a reasonable trade-off between performance and resource requirements.*

4. Results

This section presents the outcomes of the model's performance and evaluation. We begin by demonstrating the model's performance through various showcases, followed by an in-depth discussion of the evaluation metrics used to measure its success. These results provide a comprehensive understanding of the model's capabilities and limitations in the context of speech recognition for low-resource languages.

4.1. Model Demonstration and Performance Showcase

Figure 2 demonstrates how the Azerbaijani Speech Recognition model is operated in real-time with the help of the whisper small version using the Gradio model. It also includes an audio input part, where the user can either upload the recording or record a speech, and an audio output part, where the transcription of the Azerbaijani text is presented.

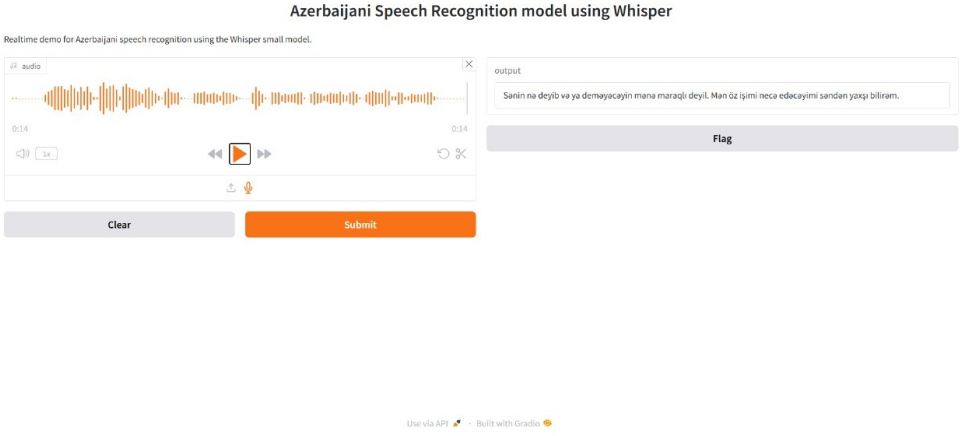

Fig. 2. Automatic Speech Recognition for Azerbaijani

In this specific example, an audio sample is played and processed by the model. The corresponding transcription shown in the output is accurately displayed in the output section. This translates to English as: What you speak and don't speak is all immaterial to me. I know how to perform my duties better than you."

I tested the transcription by speaking in real-time while pressing the microphone button and evaluating that, and the model performed quite well, returning a near-perfect accuracy. This shows actual practical applications of the model in detecting voice with unequivocal clarity.

In the paper, we cited the Google Colab notebook that was used for the Azerbaijani Speech Recognition model. This model can also be accessed in the Colab environment, and you can execute the code yourself. When executed, a temporary link is created to check the model by uploading Azerbaijani audio or recording your own voice. The details of the Colab notebook can be obtained from the reference section of the paper. [21]

4.2. Evaluation Metrics for Model Performance

The table 3 shows the evaluation measures of the Azerbaijani Speech Recognition model in the training and validation period. To monitor the model's performance, training loss, validation loss, and word error rate (WER) are recorded in different steps of training.

Table 3. Evaluation Metrics for Model Performance

Step	Training Loss	Validation Loss	WER (%)
100	0.222700	0.789307	56.36
200	0.003000	0.869487	53.94
300	0.001000	0.913662	53.33
400	0.000500	0.948102	52.73

- **Training Loss:** The training loss decreases progressively as time goes on, which is an indication that the model is learning from the training set.
- **Validation Loss:** Although there is some degree of improvement on both validation and loss, the latter stays within acceptable bounds, revealing the model capacity regarding generalization.
- **WER:** The WER at each step decreases across the steps, indicating an increase in transcript accuracy, with the lowest WER being 52.73% achieved at 400 steps.

For the purpose of this experiment, the choice of step size was made to yield a broad range of training time but still retain the effectiveness of the gradient in assessing the performance of the model given the constraint of the available computational power. However, to get an even better result, the step size could be increased. Thus, the optimization of the model would be carried out to a greater extent and, perhaps, would produce a WER of less than 10.

5. Discussion and Future Work

The performance of the Whisper small model has been found promising for speech recognition in a low-resource language after fine-tuning on an Azerbaijani speech dataset. Even with a data sample of roughly one hour, the model was quite effective during our demonstrations, with some minor omissions occasionally making it through the transcription process. These results reveal that the model possesses a robust ability to learn and function on lower amounts of data while maintaining a high degree of performance.

5.1. Discussion

The Whisper small model performance test for Low-resource language reveals that fine-tuning with the Azerbaijani speech dataset has promising results for speech recognition. However, while using the data of about one hour, the model records an impressive performance during the demonstrations only. In a few cases, some problems of transcription were reflected in the text, but they are not critical, and the overall level of performance is still satisfactory. Thus, the main idea that the model is able to learn from comparatively small volumes of materials and demonstrate acceptable reliability at the same time was confirmed.

In our case, we established that the Whisper model performance was accurate when used in a real-time Gradio interface. I found that some transcription errors occur, but in general, the performance was rather good, and thus, applying the model would be effective for the purpose. Although the analysis is based on a dataset of roughly one

hour of audio, we were able to create a tolerable set of representations to fine-tune. A small set of test speakers meant that the degree of generalization of the model over different mannerisms was somewhat impaired, and the experimental system proved to be somewhat brittle in different use cases, indicating that increasing the available training data might improve robustness.

From this, we noticed that with fewer inputs, the Whisper small model was performing well enough. It demonstrates good robustness with low-resource data, though there are some issues with outliers or changes in input data, for instance, voice variations.

5.2. Future Work

Thus, as an enhancement of the presented model, we intend to receive more Azerbaijani audio samples to increase the learning data set and improve the model. We shall also try to use bigger training steps and better hyperparameters in order to yield improved outcomes. Furthermore, it will be interesting to investigate transcription errors in greater detail and to modify the model for Azerbaijani peculiarities.

Future work will also include training on Azerbaijani-specific language models of the Whisper system for further post-processing. Further, we also intend to use the proposed model in realistic use cases such as voice assistant applications and call centers and potentially expand the database, which can then be incorporated into the Azerbaijani model and then scaled to other Low-Resource languages.

6. Conclusion

In this paper, we have presented the model Whisper for Azerbaijani speech recognition and successfully fine-tuned this model with a single hour of audio data in real-time. The model indeed had a level of success with high accuracy rates in most of the examples while at times having small transcription mistakes. These errors were mainly because of the small dataset, but otherwise, at the given resources, the model did quite well.

Despite this, the made model can be considered rather successful; thus, enlarging the data sample in the future will contribute to an increase in the model's sturdiness and effectiveness. Some aspects suggesting the improvement of the model's performance include the use of additional data in training the model, elongation of the training step, analysis of errors, and the possibility of making transcription mistakes that are typical for the Azerbaijani language. Furthermore, it is also added to integrate Azerbaijani-specific language models for post-processing in order to boost the contextual output.

The Whisper model indicates good potential for expansion to other fields, such as voice-controlled helpers, contact centers, and audiobook dictation. Increasing the size of the linguistic training data, as well as further work on applying this for multi-lingual transfer learning, could also increase the model's efficacy and applicability.

Thus, the work offers a strong starting point for constructing the linguistic ASR to Azerbaijani with a distinct direction of further augmentation and refinement to improve the model's predictive potential to integrate into practical tendencies.

Acknowledgments

We would like to express our sincere gratitude to Professors Cecilia Zanni-Merk and Stéphane Genaud for their valuable guidance and support throughout this research.

References

[1] Orken, Mamyrbayev, Alimhan, Keylan, Amirgaliyev, Beibut, Zhumazhanov, Bagashar, Mussayeva, Dinara, Gusmanova, Farida. (2020) "Multimodal systems for speech recognition." *International Journal of Mobile Communications* **18**: 314. doi: 10.1504/IJMC.2020.107097.

[2] Mussakhojayeva, Saida, Dauletbek, Kaisar, Yeshpanov, Rustem, Varol, Huseyin Atakan. (2023) "Multilingual Speech Recognition for Turkic Languages." *Information* **14** (2): 74. doi: 10.3390/info14020074. URL: https://www.mdpi.com/2078-2489/14/2/74.

[3] Tombaloğlu, B., Erdem, H. (2020) "Deep Learning Based Automatic Speech Recognition for Turkish. Sakarya University." *Journal of Science* **10.16984**: saufenbilder.711888. doi: 10.16984/saufenbilder.711888.

[4] Taşar, D. E., Koruyan, K., Çılgın, C. (2024) "Transformer-based Turkish automatic speech recognition." *Acta Infologica* **Advance Online Publication**. doi: 10.26650/acin.1338604. Available at: https://doi.org/10.26650/acin.1338604.

[5] Klakow, D., Peters, J. (2002) "Testing the correlation of word error rate and perplexity." *Speech Communication* **38** (1-2): 19–28. doi: 10.1016/S0167-6393(01)00041-3.

[6] Schneider, S., Baevski, A., Collobert, R., Auli, M. (2019) "Wav2Vec: Unsupervised Pre-training for Speech Recognition." Available at: https://arxiv.org/pdf/1904.05862.

[7] Taşar, D. E., Koruyan, K., Çılgın, C. (2024) "Transformer-based Turkish automatic speech recognition." *Acta Infologica*. Advance Online Publication. 10.26650/acin.1338604. Available at: https://doi.org/10.26650/acin.1338604.

[8] Kimanuka, U., BUYUK, O. (2018) "Turkish Speech Recognition Based On Deep Neural Networks." *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi* **22**: 10. 10.19113/sdufbed.12798.

[9] Valizada, A., Jafarova, S., Sultanov, E., Rustamov, S. (2021) "Development and Evaluation of Speech Synthesis System Based on Deep Learning Models." *Symmetry* **13**: 819. 10.3390/sym13050819.

[10] Abbasov, A., Fatullayev, A. (2007) "Forming the set of recognition units for the speech recognition system for the Azerbaijani language." *Journal Name* **Volume**: pages.

[11] Mehraliyev, A., Haziyeva, K., Almasova, M. (2022) "Speaker Verification Systems for Azerbaijani Language." *Journal Name* **Volume**: pages.

[12] Morris, E. (2021) "Automatic Speech Recognition for Low-Resource and Morphologically Complex Languages." *Thesis*, Rochester Institute of Technology.

[13] Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2018) "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models." *In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778. DOI: https://doi.org/10.1109/ICASSP.2018.8462105.

[14] Alguliyev, R., and Sukhostat, L. (2023) "Issues of Speech Technologies Application in the Azerbaijani Language." *Proceedings of the Conference on Information and Communication Technology Problems in the Azerbaijani Language*, Institute of Information Technology, Baku, Azerbaijan. Available: mailto:r.alguliev@gmail.com & mailto:lsuhostat@hotmail.com.

[15] Yi, C., Wang, J., Cheng, N., Zhou, S., and Xu, B. (2021) "Transfer Ability of Monolingual Wav2vec2.0 for Low-resource Speech Recognition." *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1-6. doi: https://doi.org/10.1109/IJCNN52387.2021.9533587.

[16] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019) "Wav2Vec: Unsupervised Pre-training for Speech Recognition." *arXiv*, https://arxiv.org/abs/1904.05862.

[17] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020) "Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." *arXiv*, https://arxiv.org/abs/2006.11477.

[18] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020) "Conformer: Convolution-augmented Transformer for Speech Recognition." *arXiv*, https://arxiv.org/abs/2005.08100.

[19] Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. (2022) "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm." *arXiv*, https://arxiv.org/abs/2110.05208.

[20] Abramenkov, A., Petuhova, N., and Farkhadov, M. (2012) "System of speech access in the Azerbaijani language to objects of electronic map." In *2012 IV International Conference "Problems of Cybernetics and Informatics" (PCI)*, pp. 1-2, doi: 10.1109/ICPCI.2012.6486263.

[21] Shikhaliyev, H. and Rustamova, M. (2024) "Colab Notebook for Speech-to-Text Recognition." Available at: https://colab.research.google.com/drive/1cEJyvlJejj7noT3eUwkri-0hcQlf5em6?usp=sharing.

[22] Pratama, R., Amrullah, A. (2024). "Analysis of Whisper Automatic Speech Recognition Performance on Low Resource Language." *Jurnal Pilar Nusa Mandiri*, 20, 1-8. 10.33480/pilar.v20i1.4633.

[23] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2022). "Robust Speech Recognition via Large-Scale Weak Supervision." *OpenAI*. Retrieved from https://cdn.openai.com/papers/whisper.pdf.

[24] Meng, L., Kang, J., Wang, Y., Jin, Z., Wu, X., Liu, X., Meng, H. (2024). "Empowering Whisper as a Joint Multi-Talker and Target-Talker Speech Recognition System." *arXiv*. Retrieved from https://arxiv.org/abs/2407.09817.

[25] Polat, H., Turan, A. K., Koçak, C., Ulaş, H. B. (2024). Implementation of a Whisper Architecture-Based Turkish Automatic Speech Recognition (ASR) System and Evaluation of the Effect of Fine-Tuning with a Low-Rank Adaptation (LoRA) Adapter on Its Performance. *Electronics, 13*(21), 4227. https://doi.org/10.3390/electronics13214227.

[26] OpenAI. (n.d.). Whisper: OpenAI's Speech Recognition Model. Retrieved from https://openai.com/index/whisper/.

[27] Oyucu, S. (2023). Comparing The Fine-Tuning and Performance of Whisper Pre-Trained Models for Turkish Speech Recognition Task. In *Proceedings of the ISMSIT 2023 Conference* (pp. 1-4). IEEE. 10.1109/ISMSIT58785.2023.10304891.