Development of a real-time speech recognition system for the Azerbaijani language

Alakbar Valizada
Azerbaijan Technical University, Baku, Azerbaijan
alakbar.valizada@aztu.edu.az

***Abstract* — Real-time automatic speech recognition (ASR) systems have become increasingly important in today's fast-paced and interconnected world. This paper presents a real-time Automatic Speech Recognition (ASR) system for the Azerbaijani language, using a hybrid Hidden Markov Model/Deep Neural Network (HMM/DNN) model for acoustic modeling and a syllable-based n-gram model for language modeling. The ASR system is designed to handle streaming speech data transmitted via WebSockets, enabling efficient real-time recognition of speech in various applications, such as virtual assistants and transcription services. The proposed system leverages the Kaldi and SRILM toolkits for model training, and incorporates a WebSocket-based communication framework to facilitate seamless data transfer between the client and server. Experiments show that the syllable-based subword modeling approach is particularly effective for Azerbaijani, an agglutinative language, and the real-time ASR system delivers promising results in terms of accuracy and performance.**

DOI: 10.25045/AzICT&ICTAz.2023.10
Keywords — ASR; HMM/DNN; n-gram; WebSocket; Kaldi; agglutinative languages; real-time ASR

I. INTRODUCTION

Automatic speech recognition (ASR) has become a critical technology in recent years, with applications in a wide range of fields, such as transcription services, virtual assistants, and human-computer interaction. As the demand for real-time speech recognition grows, it is essential to develop systems capable of processing and transcribing spoken language with minimal latency. This is especially crucial for under-resourced languages, where the availability of ASR systems is limited. One such language is Azerbaijani, an agglutinative language with a rich morphological structure, which presents unique challenges for ASR systems.

Previous research in Azerbaijani ASR has mainly focused on offline speech recognition systems for specific applications, such as emergency call centers [1] and taxi call service systems [2]. While these offline systems have made significant progress, the demand for real-time ASR systems is rapidly increasing, highlighting the need for efficient and accurate real-time speech recognition for Azerbaijani.

This paper introduces a real-time ASR system for the Azerbaijani language,

specifically designed to handle streaming speech data. The proposed system utilizes a WebSocket-based communication framework, which allows efficient transfer of speech data in real-time between clients and the recognition server. The system is built upon a hybrid HMM and time-delay neural network (TDNN) architecture for acoustic modeling, and employs a syllable-based n-gram model for language modeling. This combination allows the system to better handle the complexities of the Azerbaijani language, which often features a large number of unique word forms due to its agglutinative nature.

The implementation of the system relies on the Kaldi [3] and SRILM toolkits [4] for model training and evaluation. The paper presents the experimental setup, results, and performance metrics of the system, highlighting its effectiveness in real-time speech recognition for the Azerbaijani language.

The remainder of this paper is organized as follows: Section II provides an overview of related work in ASR and WebSocket-based communication. Section III details the experimental setup and results, including data preparation, model training, and evaluation metrics. Finally, Section IV concludes the paper and discusses future work.

II. SYSTEM OVERVIEW

A. *Architecture*

The proposed system consists of the following components:

1) *Client-side application:* Records user speech and streams audio data to the server using WebSocket.
2) *Server-side application:* Receives audio data, performs speech recognition using a hybrid HMM/DNN model, and sends recognition results back to the client.
3) *Acoustic model:* A hybrid HMM/DNN model trained using the Kaldi toolkit.
4) *Language model:* An n-gram language model created using the SRILM toolkit.

B. *WebSocket Implementation*

WebSocket is a protocol that enables bidirectional, low-latency communication between a client and a server over a single, long-lived connection (Fig. 1). This protocol is particularly suitable for real-time applications such as speech recognition, where data needs to be streamed continuously from the client to the server, and recognition results need to be returned promptly. In this section, we will discuss the server-side and client-side implementations of WebSocket for the proposed real-time speech recognition system for Azerbaijani.

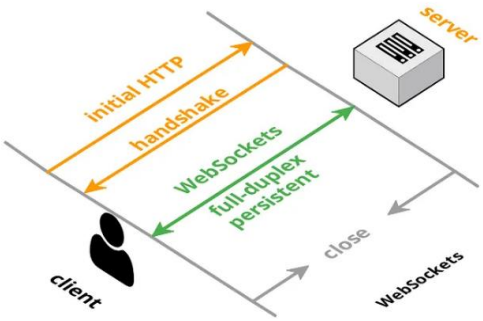

Fig.1. High-level overview of a WebSocket communication

The proposed system consists of the following components:

1) *Client-Side Implementation*

a) Recording audio data: The client application, which can be a web or mobile application, initiates a WebSocket connection to the server by providing the server's URL and the desired connection protocol. Upon establishing the connection, the client application maintains the connection throughout the user's interaction with the system.

The client-side application records the user's speech using the device's microphone and an appropriate audio recording API. The recorded audio data is typically stored in a buffer, which is a temporary storage area in memory.

To segment the audio data into smaller chunks, the client application sets up a timer or an event-based callback mechanism that triggers at regular intervals, say every T millisecond. When the timer or event is triggered, the client application retrieves the audio data from the buffer corresponding to the specified time interval.

For example, if the buffer size is set to 16,000 samples (corresponding to 1 second of audio at a 16 kHz sampling rate), and the timer interval is set to 250 milliseconds, the client application would retrieve 4,000 samples (250 ms * 16 samples/ms) from the buffer at each interval. These samples represent an audio chunk that is ready for transmission.

b) Sending audio data: Once the audio chunk has been prepared, the client application sends it to the server over the WebSocket connection. The audio data is typically encoded into a binary format, such as PCM, which preserves the original audio information and ensures efficient transmission.

Each WebSocket message sent to the server includes the binary-encoded audio chunk and any necessary metadata, such as the client's session ID, the current audio timestamp, or a sequence number to maintain the order of the chunks. The metadata can be sent as a separate message or included within the binary message, depending on the implementation.

c) Receiving recognition results: The client application listens for incoming messages from the server, which contain the recognition results. Upon receiving a message, the client application decodes the JSON object and updates the user interface to display the recognized speech in real-time.

2) *Server-Side Implementation:* On the server side, the application receives the

incoming WebSocket messages containing the audio chunks and metadata. The server is responsible for decoding the messages, extracting the audio data, and appending it to the appropriate audio buffer associated with the client's session ID.

As the audio chunks are received in the correct order (based on the sequence number or timestamp), the server application can concatenate them to reconstruct the original audio stream. The server maintains an audio buffer for each client session, ensuring that the audio data is correctly processed for each user.

When a sufficient amount of audio data has been accumulated in the buffer, the server processes the buffered audio using the ASR engine, which performs the speech recognition task. The length of the buffered audio can be determined by a fixed duration (e.g., 1 second), or adaptively based on the system's requirements and performance.

Once the ASR engine completes the recognition task, the server prepares the recognition results as a JSON object and sends it back to the corresponding client over the WebSocket connection, allowing the client to receive the results in real-time.

This process of segmenting audio data into chunks, transmitting them over WebSocket, and server-side collection and processing ensures that the real-time speech recognition system for Azerbaijani can efficiently and accurately recognize speech in a real-time scenario.

3) Acoustic model

The hybrid HMM/DNN architecture used in this study combines the strengths of HMMs and DNNs for acoustic modeling.

a) HMM Component

The HMM component is responsible for modeling the temporal structure of speech using a sequence of hidden states, $= q_1, q_2, \dots, q_T$, and observations, $O = o_1, o_2, \dots, o_T$. O observations represent the sequence of T acoustic feature vectors extracted from the speech signal at regular time intervals. Q hidden states represent the sequence of T underlying phonetic units or triphones corresponding to the observed acoustic features. The hidden states are not directly observable, and the goal of the HMM-based ASR system is to determine the most likely sequence of hidden states given the observed acoustic features [5].

The HMM is defined by the following parameters:

- State transition probabilities (A): $a_{ij} = P(q_t = j | q_{t-1} = i)$ represents the probability of transitioning from state i to state j.
- Emission probabilities (B): $b_j(o_t) = P(o_t | q_t = j)$ represents the probability of observing feature vector ot at time t, given the state j.
- Initial state probabilities (π): $\pi_i = P(q_1 = i)$, represents the probability of starting in state i.

The likelihood of a given observation sequence, O, can be calculated using the forward algorithm, which computes the forward probability, $\alpha_t(i)$, as follows:

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, q_t = i | \lambda), \quad (1)$$

where $\lambda = (A, B, \pi)$ are the HMM parameters.

b) DNN Component

While HMMs are effective in modeling the temporal aspects of speech, they are limited in their ability to model the complex relationships between the acoustic features and the phonetic units. DNNs have proven to be highly effective in modeling these complex relationships due to their ability to learn hierarchical representations of the input features [6].

In the hybrid HMM/DNN architecture, the DNN is used to model the observation likelihoods $b_j(o_t)$ for each HMM state j. Specifically, the DNN is trained to predict the posterior probabilities $P(q_t=j \mid o_t)$ of the HMM states given the acoustic features o_t. Once trained, the DNN can be used to estimate the observation likelihoods for the HMM by applying Bayes' rule:

$$b_j(o_t) = P(o_t|q_t = j) = P(q_t = j|o_t) * P(o_t)/P(q_t = j), \quad (2)$$

where $P(o_t)$ and $P(q_t=j)$ are the priors for the observations and states, respectively.

By combining the strengths of HMMs and DNNs, the hybrid HMM/DNN architecture allows for more accurate modeling of the complex relationships between the acoustic features and phonetic units in speech signals, leading to improved recognition performance.

For the training of the AM, we used a time-delay neural network (TDNN) architecture [7] with maximum mutual information (MMI) sequence-level objective function [8]. TDNNs are a type of deep neural network (DNN) that have been widely adopted for speech recognition tasks due to their ability to model both short-term and long-term temporal dependencies in the input data.

4) Language Modeling

The n-gram language model computes the probability of a sequence of words, $W = w_1, w_2, \dots, w_N$, based on the joint probability of the words. Using the chain rule of probability, we can write:

$$P(W) = P(w_1, w_2, \dots, w_N) = \prod_{i=1}^{N} P(w_1|w_{i-1}, w_{i-2}, \dots, w_1) \quad (3)$$

for an n-gram model, the probability of a word depends only on the previous n-1 words:

$$P(w_i|w_{i-1}, w_{i-2}, \dots, w_1) \approx P(w_i|w_{i-1}, w_{i-2}, \dots, w_{i-n+1}), \quad (4)$$

III. Experimental Setup and Results

A. Data Preparation

1) Acoustic Model Data

For the acoustic model, audio data were collected from two main sources. The first source was a Telegram bot, which allowed users to contribute speech samples through a crowdsourcing approach. Users were prompted to record and submit phrases from a predefined list, ensuring the diversity and coverage of the collected data. The second source of audio data was from YouTube videos. Videos containing Azerbaijani speech were selected, and the audio tracks were extracted for further processing. To create a balanced dataset, both sources of data were combined, segmented, and annotated with corresponding transcriptions. The resulting dataset was then split into training, validation, and testing sets to enable the training and evaluation of the HMM/TDNN acoustic model.

2) Language Model Data

The language model was trained using text data collected from two main sources: Azerbaijani Wikipedia [9] and the AzerTag news portal (azertag.az) [10]. These sources were chosen due to their extensive coverage of various topics, which ensured the language model would be exposed to a wide range of vocabulary and linguistic structures. The collected text data were cleaned, tokenized, and processed to generate an n-gram language model using the SRILM toolkit [4].

A significant modification was made to the language modeling approach by adopting a syllable-based subword model instead of a traditional word-based model. This approach was chosen due to the agglutinative nature of the Azerbaijani language, which often results in a large number of unique word forms. By using a syllable-based subword model, the system can better handle out-of-vocabulary words and achieve more accurate recognition results [11].

B. Acoustic and Language Model Training

The experiments described in this paper were carried out using both the Kaldi and SRILM toolkits. The configuration of the Kaldi speech recognition system is based on the Kaldi chain recipe (s5_r2) of the TED-LIUM corpus, which integrates a hybrid DNN/HMM system. The "nnet3 chain" implementation was utilized for training the acoustic model (AM), employing a TDNN architecture with a maximum mutual information (MMI) sequence-level objective function. The context-dependent states, acquired through the forced alignments of a GMM/DNN baseline system, served as targets for DNN training. Additionally, online i-vectors were used as input to the TDNN for improved speaker adaptation.

For the language model training, the SRILM toolkit was employed to train various language models (LMs). Based on our prior experience [1], modified Kneser-Ney smoothing has proven to be the most effective method for handling out-of-vocabulary (OOV) words and large text corpora.

C. Evaluation Metrics

The system was evaluated using the Word Error Rate (WER) and real-time factor (RTF) metrics. The WER measures (based on the Levenshtein distance?) the recognition accuracy, while the RTF indicates the efficiency and speed of a ASR system.

D. Results

The proposed system achieved a WER of 5.59%, which is competitive with state-of-the-art ASR systems for Azerbaijani. The system's performance can be further improved by incorporating additional data sources and refining the acoustic and language models.

The real-time factor (RTF) of the proposed system was found to be 0.165, indicating that the system processes the input speech efficiently, making it suitable for real-time applications. The use of WebSockets for streaming speech data and the optimized hybrid HMM/DNN model contribute to the system's efficiency.

IV. CONCLUSION AND FUTURE WORK

This paper presents a real-time speech recognition system for the Azerbaijani language, leveraging WebSockets for data streaming and a hybrid HMM/DNN model for ASR. The proposed system demonstrates promising results in terms of both accuracy and efficiency, making it a valuable contribution to the field of ASR for Azerbaijani. The Kaldi toolkit and SRILM toolkit were used for training the system, providing a solid foundation for further research and development.

Future work can focus on improving the system's performance by incorporating additional data sources, refining the acoustic and language models, and exploring alternative ASR architectures such as end-to-end neural models. Additionally, the system can be extended to support multilingual and code-switching scenarios, broadening its applicability and usefulness in real-world applications.

In conclusion, the proposed real-time speech recognition system for Azerbaijani offers a robust and efficient solution for ASR tasks, paving the way for a range of applications such as transcription services, voice assistants, and real-time translation.

ACKNOWLEDGMENT

I would like to thank Dr. Samir Rustamov for discussions and suggestions. This research experiments have been implemented in the Center for Data Analytics Research (CeDAR) at the ADA University/

REFERENCES

[1] A. Valizada, N. Akhundova, S. Rustamov, "Development of Speech Recognition Systems in Emergency Call Centers." Symmetry 2021, 13, 634, doi: 10.3390/sym13040634

[2] S. Rustamov, N. Akhundova, A. Valizada. "Automatic Speech Recognition in Taxi Call Service Systems," In Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Emerging Technologies in Computing; Springer: Cham, Switzerland, 2019; pp. 243–253.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, J. Silovsky: "The Kaldi speech recognition toolkit," In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society (2011)

[4] A. Stolcke (2002), SRILM - An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989, doi: 10.1109/5.18626.

[6] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[7] V. Peddinti, D. Povey, S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in: Sixteenth Annual Conference of the International Speech Communication Association (2015)

[8] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, S. Khudanpur, "Purely sequence trained neural networks for asr based on lattice free mmi," (author's manuscript). Tech. rep., The Johns Hopkins University Baltimore United States (2016), doi: 10.21437/Interspeech.2016-595

[9] "azwiki dump," 2023. [Online]. Available: https://dumps.wikimedia.org/azwiki/20230320/

[10] "The Azerbaijan State News Agency," 2022. [Online]. Available: https://azertag.az/

[11] A. Valizada, "Subword Speech Recognition for Agglutinative Languages," 2021 IEEE 15th International Conference on Application of Information and Communication Technologies(AICT), Baku, Azerbaijan, 2021, pp. 1-6, doi: 10.1109/AICT52784.2021.9620466.