ISSN: 2093-5374 / E-ISSN: 2093-5382

Development of Speech Recognition in Wireless Mobile Networks for An Intelligent Learning System in Language Education

Nargis Kurbanazarova[1*], Dilnavoz Shavkidinova[2], Murodilla Khaydarov[3], Nazmiya Mukhitdinova[4], Khuriyat Khudoymurodova[5], Dilfuza Toshniyozova[6], Nodir Karimov[7], and Rakhima Alimova[8]

[1*]Termez State University, Uzbekistan. qurbonazarovan@tersu.uz, https://orcid.org/0000-0003-1384-8302

[2]"TIIAME" National Research University, Uzbekistan. dilnavoz.shavqidinova@gmail.com, https://orcid.org/0009-0002-2778-1030

[3]University of Geological Sciences, Uzbekistan. xaydarov@mail.ru, https://orcid.org/0000-0003-4302-4080

[4]Samarkand State University named after Sharof Rashidov, Uzbekistan. muxitdinovanazmiya@gmail.com, https://orcid.org/0000-0001-7108-7244

[5]Termez State University, Uzbekistan. hudoymurodovaxurriyat@gmail.com, https://orcid.org/0000-0003-4768-8503

[6]Samarkand State University named after Sharof Rashidov, Uzbekistan. toshniyozovadilfuza18@gmail.com, https://orcid.org/0009-0001-0916-1519

[7]Tashkent State University of Oriental Studies, Uzbekistan. nodir-karimov@list.ru, https://orcid.org/0000-0001-5127-8713

[8]Tashkent State University of Oriental Studies, Uzbekistan. alimova.rahima@gmail.com, https://orcid.org/0009-0009-9434-9601

Received: May 22, 2024; Revised: July 26, 2024; Accepted: August 22, 2024; Published: September 30, 2024

Abstract

Communication has been crucial to human existence, society, and globalization for millennia. Speech Recognition (SR) technologies include biometric evaluation, security, safety, medical care, and smart cities. Most research has primarily focused on English; others neglect other lower-asset dialects like Uzbek, neglecting its research unaddressed. This study examines the efficacy of peer and ASR response in wireless mobile networks-assisted pronouncing training. This study proposes a Deep Neural Network (DNN) and Hidden Markov Method (HMM) based ASR model to develop a voice recognition system utilizing a combination of Connected Time-based Categorization (CTC)-attending networks for the Uzbek words and their variants. The suggested method diminishes training duration and enhances SR precision by efficiently employing the CTC goal function in attentiveness modeling. The research assessed the results of both linguistic experts and native speakers on the Uzbek database, which was compiled for this research. The data were gathered

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 3 (September), pp. 298-311. DOI: *10.58346/JOWUA.2024.I3.020*

*Corresponding author: Termez State University, Uzbekistan.

through a pronunciation assessment and a discussion. The participant was further instructed in the classroom. Test outcomes indicate that the suggested method attained a word error rate of 13.1%, utilizing 210 hours of records as a learning dataset for the Uzbek dialect. The proposed technique can significantly enhance students' pronunciation qualities. It might inspire pupils to participate in pronunciation learning.

Keywords: Speech Recognition, Wireless Mobile Networks, Smart Learning, Language Education.

1 Introduction

Speech is the fundamental and innate mode of human interactions, adept at transmitting significant data swiftly and accurately (Mehrish et al., 2023) Individuals dedicate time and effort to mastering communication with various intelligent gadgets through vocal instructions. The uniqueness of language has been identified in 7,080 spoken languages globally. A living dialect is defined as having more than one individual who utilizes it as their primary language. The wide variety of languages used does not suggest an equitable distribution among people worldwide; over half of the world's populace communicates using just 24 dialects, such as Chinese, English, German, and Hindi (Ramoo et al., 2021). The most widely spoken dialects are extensively documented, facilitating the development of Artificial Intelligence (AI) (Deka et al., 2023) technologies for Texts-To-Speeches (TTS) (Kaur & Singh, 2023), Automated Speech Recognition (ASR) (Li et al., 2024). Natural Language Processing (NLP) (Kang et al., 2020), and mathematical semantics. Lesser-known dialects need more assets for particular technological advancement and study (Sulochana, 2020). Developing analogous solutions for lower-asset dialects is a formidable and significant undertaking.

ASR is a vital and dynamic field of study because of its extensive applications in safety, training, intelligent medical care, and smart cities, alongside the advancement of interfaces and computational tools that facilitate voice analysis. It synthesizes many methodologies that enable the transformation of auditory data into text, employing text comparison on the identified speech signals present in the output (Esposito et al., 2024). ASR aims to transform speech signals into textual information by establishing a reliable foundation for enhanced semantic comprehension (Verkholyak et al., 2021). ASR encompasses computer technological advances, Digital Signal Processing (DSP), artificial intelligence, research, and disciplines (Conghai et al., 2021).

There is a significant transition in ASR technologies from DNN-depended mixed simulation to End-To-End (E2E) system. Hybrid designs necessitate the distinct improvement of individual component models, including pronunciation, sound, and language modeling. E2E ASR systems simultaneously convert input speech sequences into a result token sequencing using a unified network, whereby the components of the pronunciation, sound, and language models are collectively trained inside a singular framework.

The most advanced ASR for the Uzbek language has been delivered by module Deep Neural Networks (DNN) (Oh et al., 2020) and Hidden Markovian Modeling (HMM) devices (Mor et al., 2021). The most favorable ASR outcomes on the Uzbek-speaking database information have been released, with a word error rate of 13.1% on the Uzbek-speaking database testing collection (Nematova et al., 2024). Two significant challenges arise while developing voice recognition algorithms for the Uzbek dialect:

New phrases are generated by appending various suffixes to the roots, thus expanding vocabulary length and introducing numerous Uzbek languages with minimal identified information using wireless

mobile networks. Every word is a Uzbek dialect that can be heard yet lacks a text form due to the absence of established orthographic conventions.

The research presents an ASR system utilizing the proposed framework for Uzbek speech and its variants to resolve these issues. Hybrid Connectionist Timing Categorization (CTC) was an end-to-end solution (Bansal et al., 2022). The suggested method diminishes training duration and enhances Speech Recognition (SR) precision by efficiently employing the CTC goal function in learning attention models.

The main findings of the research are as follows:

An integrated ASR system integrating E2E and HMM was suggested within a DNN acoustic modeling utilizing shared networks. This represents the inaugural effort to employ Deep-Learning (DL) methodologies in the creation of an end-to-end Uzbek speech and dialect SR structure; different cutting-edge position-embedding strategies were explored utilizing a convolution-augmented converter for Uzbek SR; multiple advanced analytical approaches were introduced to support the hypotheses from the suggested ASR structure, resulting in enhanced efficiency and total system development; The research performed trials and obtained substantial outcomes for Uzbek language voice recognition utilizing an end-to-end architecture and the proposed model using wireless mobile networks.

This work advances the creation of a Vocal-based Activity Detector (VAD) pathway for the end-to-end transformers to address the challenge of excessively long speech sections, a frequent issue in exploratory ASR systems. The developed pipeline integrates the precision with the optimum duration of the VAD features using wireless mobile networks. The primary goal is establishing a new standard, incorporating cutting-edge designs to render them available to the ASR.

2 Background

1) ASR and Training

ASR is extensively implemented to enhance students' language abilities. Research indicates that ASR enhances learners' pronunciation excellence, correctness in spoken grammatical frameworks, and proficiency in speech acts. ASR provides learners with a comfortable and pleasurable atmosphere, alleviating their speaking apprehension (Dai & Wu, 2023). The efficacy of ASR is underscored by Nickolai et al. after examining 350 empirical research on language learning systems (Nickolai et al., 2024).

A recent advancement in ASR-assisted pronouncing training is the academic focus on mobile-based dictating ASR. Unlike computer-based ASR technologies and language acquisition applications, mobile-assisted dictating ASR does not provide an evaluation score or identify wrong utterances (Jiang et al., 2023). Students must analyze the dictated word to discover pronouncing errors. Despite its apparent simplicity, wireless mobile networks-based dictating ASR feedback significantly aids learners in acquiring pronunciation skills. Zou et al. examined learners' pronouncing acquisition under three circumstances: ASR response (using the Nuance Dragon Speaking application), instructor comments, and the absence of feedback (Zou et al., 2020). The ASR response group exhibited substantial improvement, while the other two categories did not. Lai et al. discovered that learners gained advantages by utilizing Mailing services as a dictating ASR instrument in conversations (Lai & Chen, 2024). The students enhanced their drive and readiness to speak, with more than half demonstrating better oral ability. Although these investigations highlight the advantages of dictating ASR, they predominantly concentrate on independent learning. Independence for learners is undoubtedly

significant for language acquisition, yet learner engagement and cooperation are equally essential. Considering the improved connectedness facilitated by mobile technology, it is necessary to comprehend how ASR-assisted cooperative training might increase pupils' pronunciation using wireless mobile networks.

2) Collaborative Speaking Training in Virtual Aspects

In cooperative speaking education, peer communication provides valuable chances for trainees to practice specific pronunciation elements with the support of their peers (Udayakumar et al., 2023). Current studies indicate that peer response combined with ASR enhances learners' feedback and achievement in pronouncing acquisition (Doris et al., 2023). Fendji et al., discovered that students appreciated peer input and had a heightened awareness of the advantages and disadvantages of their speech (Fendji et al., 2022). In a study, English majoring undergraduates were allocated into three therapy circumstances: single system-based, collaboration automated, and single non-system-based (Llopiz-Guerra et al., 2024). Every category enhanced their speaking; however, the improvements were not statistically different (Odilov, 2024).

Evers & Chen, (2022) examined two cohorts of adult students trained with computer-based ASR with peer comments from those who utilized ASR comments only (Evers & Chen, 2022). The earlier group demonstrated superior performance compared to the other group in the clarity of read-aloud phrases and conversational dialogue. This research has shown that technology-facilitated peer contact enhances learners' emotional engagement and effectiveness in pronouncing acquisition (Shadiev & Yang, 2020).

3) Goals and Questions

A study of the pertinent literature indicates that prior research has concentrated on the ASR response of wireless mobile networks and compared it with in-person peer feedback. Given the pervasive use of dictating ASR in cell phones and students' reliance on cell phones for social networking, it is pertinent to investigate whether ASR and the advantages of "mobility" and "peer connection" might enhance pronouncing acquisition. The project examines the impact of wireless mobile networks, peer suggestions, and ASR evaluations on the speech training of EFL learners. The subsequent research questions led to this study:

- Do three wireless mobile network-assisted circumstances (i.e., peer suggestions, ASR suggestions, and a mix of both) enhance learners' pronunciation acquisition?
- Do students exhibit variations in their pronouncing learning under wireless mobile network-assisted circumstances?
- How do learners evaluate feedback from peers and ASR facilitated by wireless mobile networks?

3 Materials and Methods

1) Research Data

This study employed a mixed-methods methodology, incorporating pronouncing assessments and gathering participants' perspectives via surveys. The evaluations and surveys produced information when the discussions yielded information. The justification for the study's design was dual. Initially, it concentrated on both teaching results and learner perspectives, providing a thorough picture of the

results and emotions associated with the mobile-assisted pronouncing procedure. Secondly, student perspectives were obtained by both numerical and qualitative information, the triangulation that enabled us to rigorously detect, compare, and elucidate apparent disparities between wireless mobile networks-assisted peer assessment and ASR response.

2) ASR System for the Uzbek Language

The development of an SR structure entails the acquisition of audio words, transcription of recordings, connecting of letters to phonemes to create phoneme lexicons, phonetic translation, sound modeling, linguistic modeling, extracting of speech characteristics, and implementation of a learning method focused on loss minimizing it. Rather than depending on transcriptions of phonetics or synchronized audio recordings with letter-to-phoneme correspondence, the end-to-end SR approach is essential for these processes. A strategy involves supplying a raw audio recording to the ASR structure, which identifies audio features, correlates audio clips to protagonists, and presents the recognized symbols alongside their recognition probabilities using wireless mobile networks. The features are derived from speech files and input into a multiple-tier DNN that computes the probabilities for every character via the profound ASR methodology. A dialect system next converts these likelihoods into coherent words and sentences, finally delivered to the listeners.

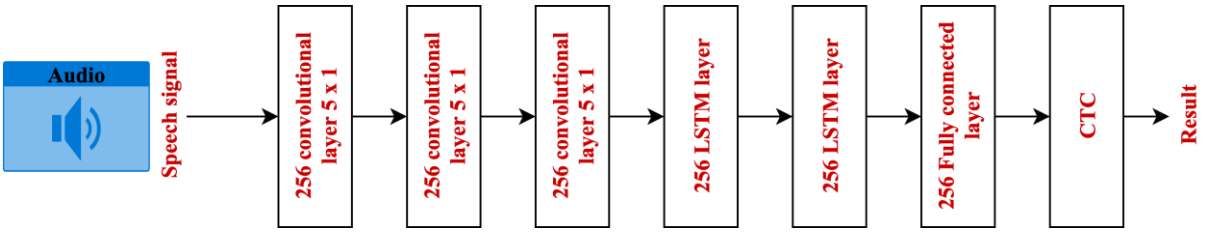


Figure 1: The architecture of the DNN model for ASR

Figure 1 illustrates the suggested network design. The vector data from the audio segment sampled every 25 ms, are inputted into the convolutional component of the networking design. The system begins with three successive convolutional layers. Subsequent layers consist of Long Short-Term Memory (LSTM) units. After the recurring layer, fully connected tiers were implemented, followed by applying the softmax activating mechanism.

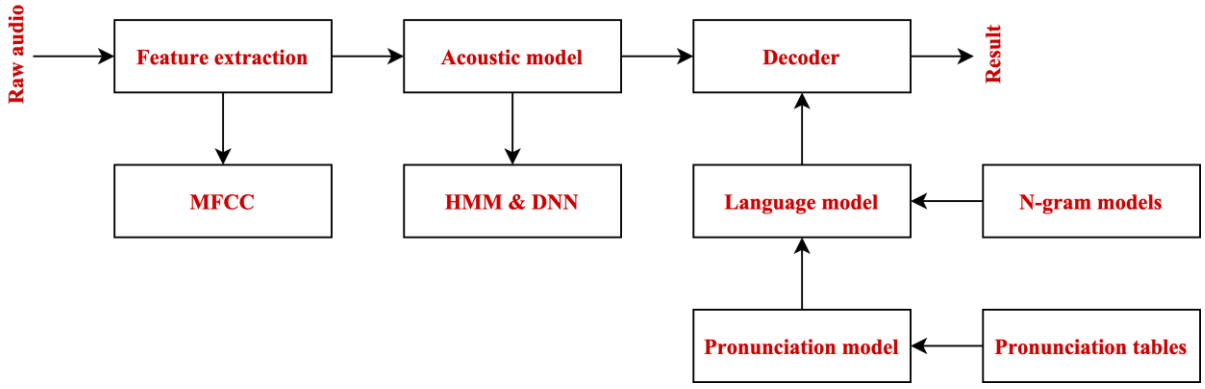


Figure 2: Overview of the Suggested ASR Model

Figure 2 illustrates the development and enhancement of ASR. ASR is a complex challenge in NLP since it involves many subdivisions, including voice segmentation, sound modeling, and linguistic

modeling, to provide predictions of patterns from distorted, unsegmented information. The advent of CTC obviates the requirement for information and facilitates end-to-end system learning for assignments, including ASR.

Extracting feature: Standardization, windows, and spectral images are employed for preparing recordings, including Mel-Frequency Cepstral Coefficients (MFCC) images or MFCCs.

Acoustic Modeling: A CTC-based system is utilized to forecast the probability probabilities of vocabulary symbols c at cycle t;

The greedy algorithm represents the most straightforward decoding technique. At every phase, the character with the highest likelihood from the time softmax outputting tier is chosen, irrespective of any semantic understanding of the transmitted content. The redundant letters are eliminated or condensed, and the vacant labels are removed;

Dialect modeling can provide data for the acoustic framework and assist in rectifying any mistakes that have arisen. Employ a beam to decode and interpret the utterance by amalgamating the acoustic system's perception with the probability of the subsequent word's occurrence inside the setting.

3) Dataset

Mozilla's exemplary open-source initiative, the Community Voice Collection 8.0, has amassed 14,000 hours of speech data in 87 dialects, rendering it a significant resource for open voice corpus development. The Uzbek dialect comprises around 227 speech recordings in the Standard Voice Collection 8.0, with just 80 hours being authenticated. A comprehensive large-vocabulary Uzbek speech collection has yet to be created and accessible. The Common Speaking and Speech-Ocean audio datasets were utilized to learn the Uzbek ASR technology employing NLP. Table 1 presents an overview of the two databases, detailing the overall word count, the overall number of distinctive phrases, the overall length, the overall number of phrases, the mean length per phrase, and the mean word count per sentence.

Table 1: Dataset Details

Dataset		Words	Unique Words	Time (hr)	Utterances	Mean time (s)	Mean word / utterances
Common voice	Training	500k	79k	132.3	118k	4.31	5
	Testing	24k	14k	12.1	8.9k	4.73	7
Common voice + speech Ocean	Training	40k	50k	78.7	62k	5.23	6
	Testing	900k	129k	210.2	183k	4.95	8

4) Modeling of Acoustic and Language

This study employed an architecture that combines a Time-DNN (TDNN) with LSTM tiers, yielding markedly superior results compared to BLSTM acoustical modeling. The TDNN architecture comprises six concealed layers, each containing 1024 concealed units. The Neural Networks (NN) underwent training utilizing the Lattice-Free Maximal Mutual Interaction (LF-MMI) methodology. The modular solution accepts characteristics of MFCC without energy and their initial and subsequent derivatives in standard 13-dimensional cepstral mean-variance normalization characteristics deprived of power using wireless mobile networks. A linear prejudiced approach was used to translate the concatenation images into 40 aspects, succeeded by the highest probability linear transformation to guarantee that each frame

included four consecutive frames. Speaker adaptation was employed with maximum probability linear regression in the characteristics area. In the proposed approach, 100k Gaussians indicate 5k variables. The research set's optimal language load and silent penalty values were 0.8 and 0.0, respectively. The acoustic modeling was developed with the Kaldi ASR toolset.

The research developed 2 n-gram language models: a large 4-gram (bLM4) training on audio recordings and a nine million-word corpus, and a bLM4 using a pool. The initial acoustic decoder employed a compact language model to generate lattices. After that, bLM4 was utilized to re-evaluate data. To ensure compatibility, the time-constrained level is used for sound modeling. The research examined sub-word modeling utilizing 128k subwords calibrated for the proposed model. These outcomes demonstrate that word encoding led to a 1.1% relative reduction in the Uzbek experimental group and a 0.7% comparative rise in WER in the Uzbek testing collection. The research decided to forgo subword analysis and utilize the advanced proposed structure with word tokenization.

5) The End-to-End Transformer

It is a contemporary architecture that entirely eradicates repetition in conventional recurring systems, opting for a self-attention process and sinusoidal positional encoding instead. Figure 3 illustrates the implementation of a transformer-based architecture for the Uzbek ASR.

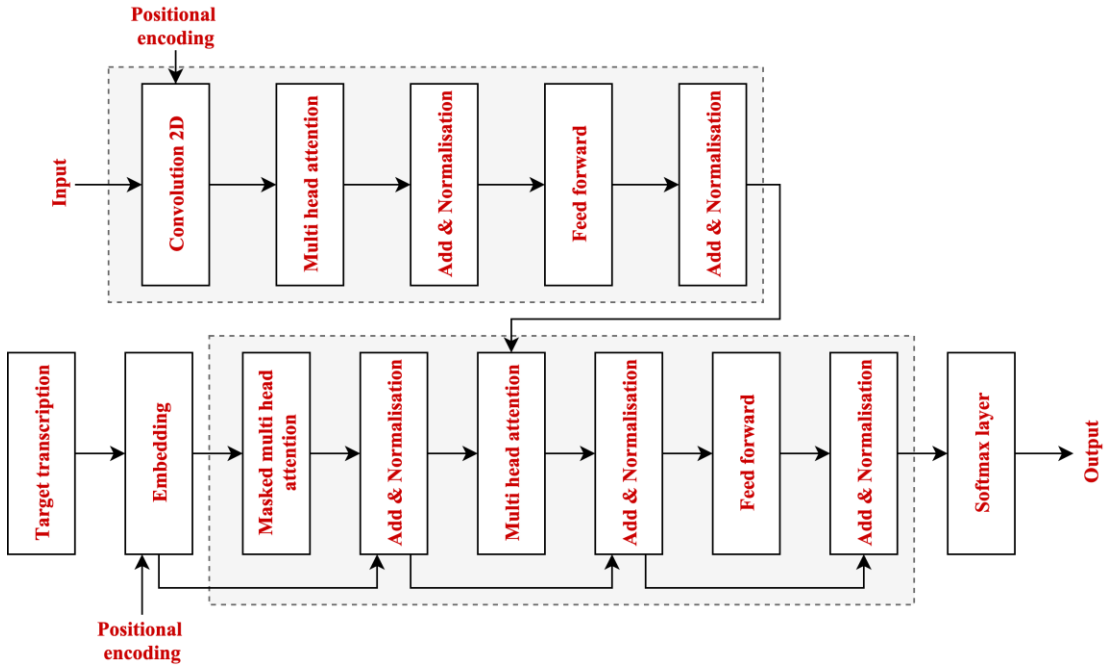


Figure 3: Ene-to-end Converter Model for ASR

It has M iterative encoding units and N iterative decoder units—diagrams showing the encoder architecture and the vector representing its hidden representations using wireless mobile networks. Log-Mel spectral images of 80 aspects and characteristic images with 83 aspects are input into the X generator. The decoding generates a single forecast sequentially. At each stage, a latent version of the encoding modeling and prior estimates from the decoding are provided.

4 Results and Discussions

A Uzbek orthographic rules dataset assessed the proposed end-to-end transformer approach against existing solutions. Standard WER and Character Error Rates (CER) metrics were utilised to determine the Uzbek database. Investigations have been undertaken to utilize DL techniques for the ASR of the Uzbek language through interconnected NNs and to identify Uzbek language instructions represented as pictures using DNN. The research comprehensively assessed the ASR technologies by comparing the traditional hybrid-based approach and E2E designs using wireless mobile networks. The trials utilized a database of the Uzbek language, encompassing both training and testing groups. The learning of the speech specimens, organized inside the suggested network design, was executed on Python Jupyter notebooks. The research was performed on High-Performance Computation (HPC) nodes with 4 Graphics processing units having 16 GB of Memory and 20 Central Processing Units.

This research will analyze the errors and relationships with language experts, native participants, the E2E converter, and DL algorithms. In a shared element across all methodologies, 50 teachers (25 male and 25 female) engaged in teaching the speech corpora. Every speaker produced a 40-minute audio clip from the readings of tales and books by different writers of contemporary Uzbek writings.

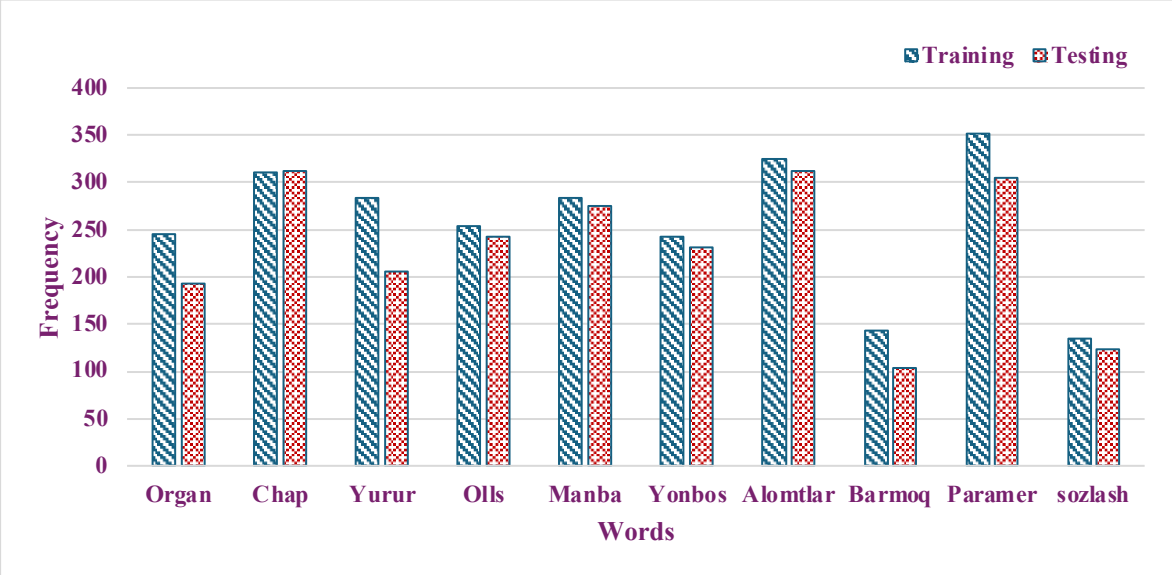


Figure 4: ASR Outcomes

The cumulative duration of the audio recordings was 24 hours. The training speech dataset had 12k phrases, among which 41k were unique. The research utilized 80% of the complete voice dataset for training, 10% for network configuration during learning, and 10% for assessment. Figure 4 illustrates the example outcomes of the suggested ASR system for Uzbek derived from the combined ASR framework. The research performed a 24-hour training period in the tests on the proposed NN design. The terms chosen for evaluation exhibited the maximum incidence in the learned dataset. ASR surpassed that of other infrequent terms. The outcome of recognizing word rate was highlighted with a red rectangle.

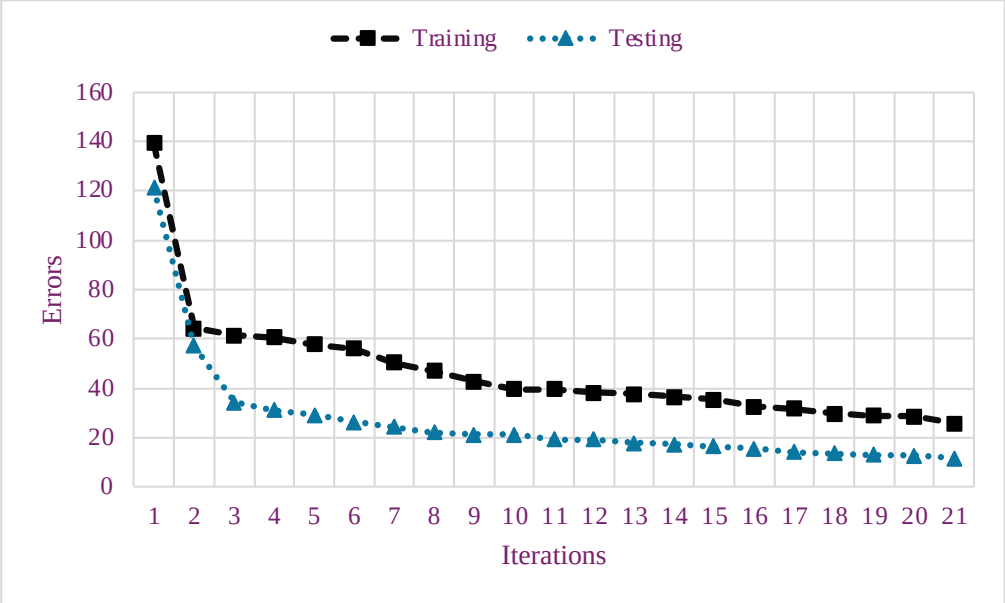


Figure 5: Error Analysis of ASR

Figure 5 illustrates that system training and testing errors were influenced by sample dimensions, speaker count, pronunciation speed in the sound file development, acoustic environment, and various other variables, including training with a restricted vocabulary or the high-precision automatic recognition of random phrases. Learning in a restricted vocabulary or high-precision automated identification of random words is contingent upon several factors, including sample dimensions, speaker quantity, pronouncing speed during audio file construction, acoustic environment, and numerous more characteristics.

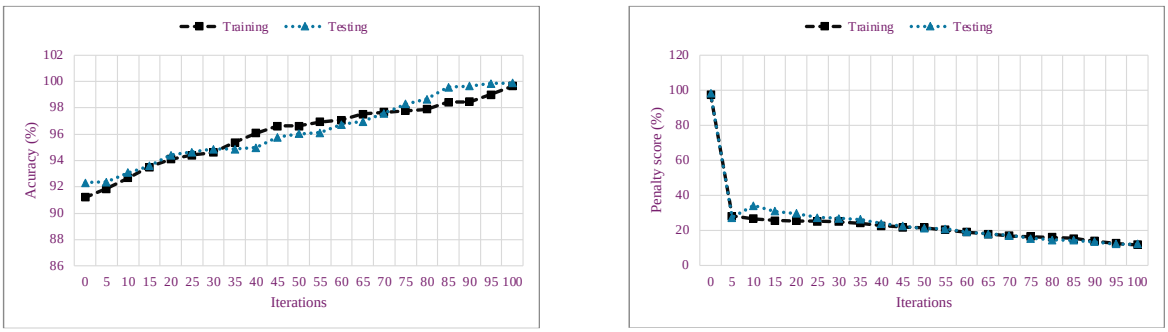


Figure 6 (a): Accuracy and 6(b): Mean Penalty Score Analysis

Figure 6(a) illustrates the effectiveness of the Uzbek voice recognition model developed using a DNN, as presented in this study. The abscissa represents the number of repetitions in learning the DNN approach. At the same time, the ordinate denotes the accuracy score of the model on both the learning and test sets. Figure 6(b) illustrates a graph depicting the mean penalty level of the network relative to the number of repetitions. The abscissa denotes the number of repetitions in learning the DNN approach. At the same time, the coordinates indicate the mean penalty score of the algorithm on both the learning and test sets.

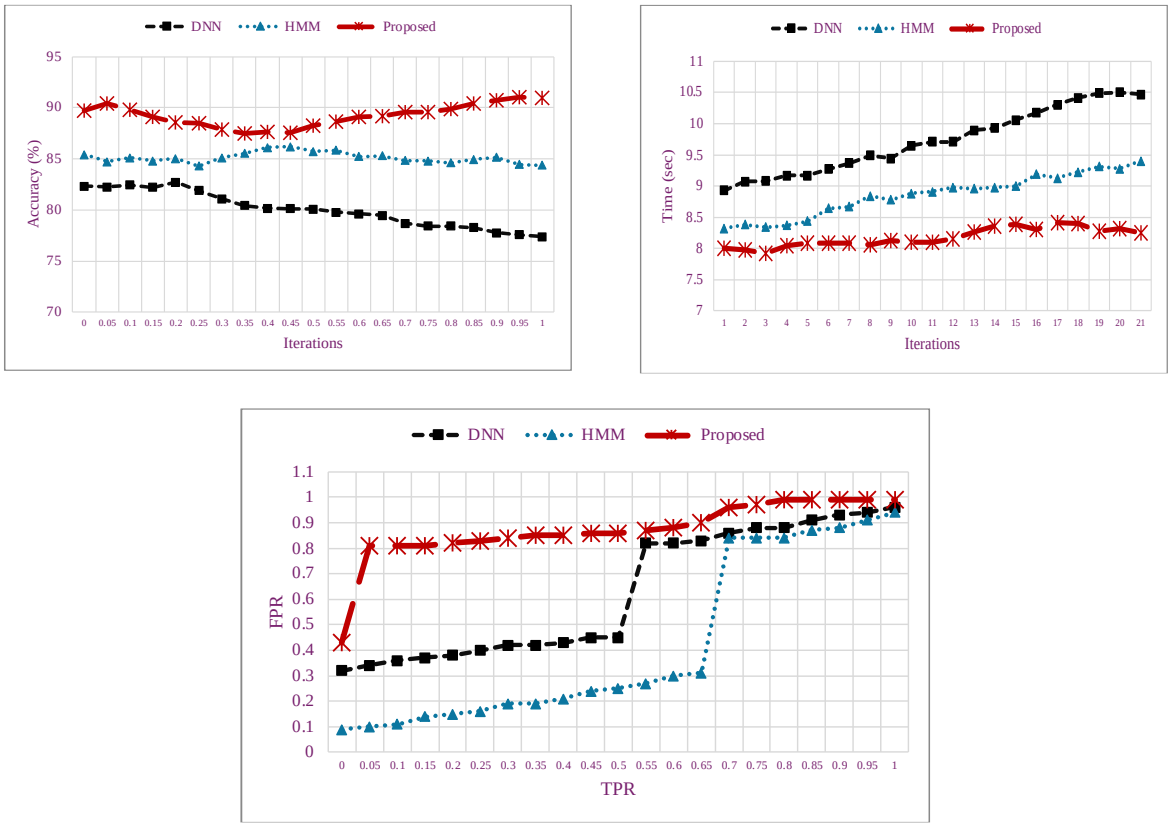


Figure 7(a): Accuracy, 7(b): Duration and 7(c): False Positive Rate Analysis of ASR

To further validate the efficacy of the proposed feature system, the research constructed the SR algorithms utilizing a DNN system, an HMM system, and a proposed network that incorporates a linear combining of speech elements and characteristics. Figure 7(a) presents a comparative efficiency analysis, while Figure 7(b) illustrates the temporal energy consumption associated with each method.

Figure 7(a) illustrates that the reliability of the proposed method presented in this research significantly surpasses that of the DNN and HMM methods. Therefore, integrating these methods leverages their benefits and enhances accuracy further. Figures 7(b) and 7(c) illustrate that the linear fusion technique, which incorporates the DL-based speech feature system and the speech characteristic approach, more effectively utilizes the data from the speech characteristic and attribute characteristics than the individual models. Despite the sequential combination of the speech characteristic modeling and the speech characteristic modeling being utilized linearly, the duration required remains equivalent to that of using either the speech characteristic modeling or the speech characteristic modeling independently. This validates the accuracy of the voice recognition system utilizing a DNN that integrates several characteristics in the Uzbek language.

5 Conclusion

This work involved the compilation of a Uzbek language voice corpus for learning the suggested ASR approach, and it provided a thorough comparison between the module DNN and HMM-based ASR and the end-to-end automatic speech recognition for the Uzbek dialect. An error study was performed to evaluate the accuracy of the sophisticated ASR algorithm compared to that of linguistics and a native

user. This research examined wireless mobile network-assisted learning environments where students got guidance from peers and ASR. The results indicate that all subgroups enhanced pronunciations for segmental precision and reliability; however, wireless mobile networks peer response appeared to have a superior training impact than ASR response. The trials demonstrated that the ASR system outperformed a native speaker considerably.

The research saw a significant resemblance between the machine mistakes and the language expert's transcription, as indicated by the WER, demonstrating a mean efficiency of 4.1% inferior to that of the linguistics in reproducing the unrefined Uzbek language. The research has built an E2E converter for Uzbek ASR and regional dialects. The suggested E2E transformer obtained superior results of 13.1% relative to existing leading models. The tests have demonstrated that in real-life ASR, fluctuation significantly impacts the efficiency of the end-to-end conversation. The suggested approach can enhance students' pronouncing abilities and engagement. This was substantiated by the pupils' performance and the marks achieved. A VAD limit was implemented to address this issue. Future research includes a longitudinal approach to examine pupil perspectives towards ASR and cooperative dynamics across various pronunciation assignments. These observations will illuminate the viability of wireless mobile networks-assisted pronunciation acquisition as a standard training regimen instead of a supplementary adjunct.

References

[1] Bansal, S., Sharan, S., & Agrawal, S. S. (2022). Study of speech recognition system based on transformer and connectionist temporal classification models for low resource language. *In International Conference on Speech and Computer, Cham: Springer International Publishing*, 56-63.

[2] Conghai, H., Qianqian, Z., & Jie, G. (2021). An artificial intelligence-based speech model for linguistics teaching. *Journal of Intelligent & Fuzzy Systems*, *40*(2), 3605-3615.

[3] Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: a mixed-methods study. *Computer Assisted Language Learning*, *36*(5-6), 861-884.

[4] Deka, C., Shrivastava, A., Nautiyal, S., & Chauhan, P. (2023). Human-Centered AI Goals for Speech Therapy Tools. *In International Conference on Computer-Human Interaction Research and Applications*, 121-136.

[5] Doris, F.G., Oscar, G.G.Z., Juan, S.T., Silvia, J.A.V., Miguel, A.S., & Ronald, M.H. (2023). An Ensemble-based Machine Learning Model for Investigating Children Interaction with Robots in Childhood Education. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *14*(1), 60-68.

[6] Esposito, M., Valente, G., Plasencia-Calaña, Y., Dumontier, M., Giordano, B. L., & Formisano, E. (2024). Bridging auditory perception and natural language processing with semantically informed deep neural networks. *Scientific Reports*, *14*(1), 20994. https://doi.org/10.1038/s41598-024-71693-9

[7] Evers, K., & Chen, S. (2022). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, *35*(8), 1869-1889.

[8] Fendji, J. L. K. E., Tala, D. C., Yenke, B. O., & Atemkeng, M. (2022). Automatic speech recognition using limited vocabulary: A survey. *Applied Artificial Intelligence*, *36*(1), 2095039. https://doi.org/10.1080/08839514.2022.2095039

[9] Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2023). Exploring the effects of automatic speech recognition technology on oral accuracy and fluency in a flipped classroom. *Journal of Computer Assisted Learning, 39*(1), 125-140.
[10] Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics, 7*(2), 139-172.
[11] Kaur, N., & Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review, 56*(7), 5837-5880.
[12] Lai, K. W. K., & Chen, H. J. H. (2024). An exploratory study on the accuracy of three speech recognition software programs for young Taiwanese EFL learners. *Interactive Learning Environments, 32*(5), 1582-1596.
[13] Li, C., Xu, W., Cohen, T., & Pakhomov, S. (2024). Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics, 150*, 104598. https://doi.org/10.1016/j.jbi.2024.104598
[14] Llopiz-Guerra, K., Daline, U.R., Ronald, M.H., Valia, L.V.M., Jadira, D.R.J.N., Karla, R.S. (2024). Importance of Environmental Education in the Context of Natural Sustainability. *Natural and Engineering Sciences, 9*(1), 57-71.
[15] Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion, 99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869
[16] Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. *Archives of Computational Methods in Engineering, 28*, 1429-1448.
[17] Nematova, S., Zinszer, B., & Jasinska, K. K. (2024). Exploring audiovisual speech perception in monolingual and bilingual children in Uzbekistan. *Journal of Experimental Child Psychology, 239*, 105808. https://doi.org/10.1016/j.jecp.2023.105808
[18] Nickolai, D., Schaefer, E., & Figueroa, P. (2024). Aggregating the evidence of automatic speech recognition research claims in CALL. *System, 121*, 103250. https://doi.org/10.1016/j.system.2024.103250.
[19] Odilov, B. A. (2024). Utilizing Deep Learning and the Internet of Things to Monitor the Health of Aquatic Ecosystems to Conserve Biodiversity. *Natural and Engineering Sciences, 9*(1), 72-83.
[20] Oh, Y. R., Park, K., & Park, J. G. (2020). Online speech recognition using multichannel parallel acoustic score computation and deep neural network (DNN)-based voice-activity detector. *Applied Sciences, 10*(12), 4091. https://doi.org/10.3390/app10124091
[21] Ramoo, D., Romani, C., & Olson, A. (2021). Lexeme and speech syllables in English and Hindi. A case for syllable structure. *Trends in South Asian Linguistics, 367*, 415. https://doi.org/10.1515/9783110753066
[22] Shadiev, R., & Yang, M. (2020). Review of studies on technology-enhanced language learning and teaching. *Sustainability, 12*(2), 524. https://doi.org/10.3390/su12020524.
[23] Sulochana, N.E. (2020). Children's Literature: A Tool to Enrich Learning in the Elementary School. *Indian Journal of Information Sources and Services, 10*(2), 48–55.
[24] Udayakumar, R., Muhammad, A.K., Sugumar, R., & Elankavi, R. (2023). Assessing Learning Behaviors Using Gaussian Hybrid Fuzzy Clustering (GHFC) in Special Education Classrooms. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14*(1), 118-125.
[25] Verkholyak, O., Dvoynikova, A., & Karpov, A. (2021). A Bimodal Approach for Speech Emotion Recognition using Audio and Text. *Journal of Internet Services and Information Security, 11*(1), 80-96.
[26] Zou, B., Liviero, S., Hao, M., & Wei, C. (2020). Artificial intelligence technology for EAP speaking skills: Student perceptions of opportunities and challenges. *Technology and the Psychology of Second Language Learners and Users*, 433-463.

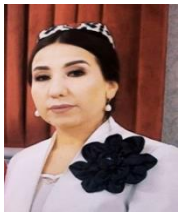
Authors Biography

Nargis Kurbanazarova, an associate professor at Termiz State University, is a PhD holder. She has published around 30 articles, a monograph, and a study guide. In 2021, Area of interest would be to bring technology to the community of education.

Dilnavoz Shavkidinova, an associate professor at TIIAME National Research University, holds a PhD in International Relations and Political Science from Sookmyung Women's University. She has degrees in English Philology, Educational Psychology, and International PR and Marketing. Her research interests include English language, IT, economics, international relations, and education.

Murodilla Khaydarov, Doctor of History and Professor, has published over 100 scientific articles and authored 5 textbooks, 7 training manuals, and 1 monograph. He defended his candidate's thesis on Soviet centralization in Turkestan and became an associate professor in 2003. He has also co-developed 3 educational programs on the history of Uzbekistan.

Nazmiya Mukhitdinova, Doctor of Philological Sciences and Professor at Samarkand State University, specializes in the poetic development of Uzbek literature from the 18th-19th centuries. She has published rare divans of classical poets and authored 3 teaching manuals, 1 textbook, and nearly 100 local and international publications.



Khuriyat Khudoymurodova, is an associate professor at the Department of Uzbek Literature since October 2022. She earned her PhD in 2020. Author of 5 books and over 100 articles, her research has been featured in prestigious international journals and conferences. In 2024, she delivered a lecture on Uzbek literature at Malaysia's University of Science.

Dilfuza Toshniyozova, is a lecturer of the classical department of Samarkand State University. She is the author of more than 40 articles on classical topics. She has found the poems of Syed Muhammed Khaifa Yuldosh ibn Davlat, whose work was unknown until now, and published them as “Devon” under the editorship of literary scholar Muhiddinov Muslihiddin.

Nodir Karimov, is an associate professor at Tashkent State University of Oriental Studies and holds a PhD. He has numerous publications and contributions to the academic community. As a recognized expert in his field, Karimov actively participates in national and international conferences. Expertized in bringing technology to the community of education.

Rakhima Alimova, is a PhD holder. She is the author of more than 100 scientific works. Among them, there are 1 textbook, 10 study guides, 2 monographs and articles in prestigious local and foreign magazines. Expertized in bringing technology to the community of education.