PROBLEMS AND SOLUTIONS IN DEVELOPING AN STT SYSTEM BASED ON UZBEK DIALECTS

Roza Urinbaeva Toliboy qizi
National University of Uzbekistan PhD Student,
Department of Computational Linguistics
and Applied Linguistics
orinboyevaroza@gmail.com
https://orcid.org/0009-0001-0442-9686

Abstract: This article discusses the main challenges and solutions in developing a Speech-to-Text (STT) system based on Uzbek language dialects. Within the framework of the study, a linguistic database covering the dialects of 12 regions of Uzbekistan was created, and dialectal speech samples were collected and scientifically annotated. The paper analyzes the impact of phonetic and pronunciation variability specific to dialects on automatic speech recognition systems and proposes approaches for improving STT models using multi-dialectal data. The research findings contribute to the development of modern speech technologies for the Uzbek language and the enrichment of national linguistic resources.

Keywords: Uzbek language, dialects, STT system, speech recognition, dialectal corpus, linguistic database, artificial intelligence.

INTRODUCTION. In recent years, the rapid advancement of artificial intelligence and speech technologies has significantly increased the demand for systems capable of enabling natural interaction between humans and computers. Speech-to-Text (STT) technologies, which automatically convert spoken language into written text, are now widely applied across various domains, including education, media, digital services, and e-governance. However, most existing systems are primarily designed for standard (literary) languages and demonstrate limited capacity to effectively process dialectal speech.

The Uzbek language is characterized by considerable regional diversity, encompassing a wide range of dialects with distinct phonetic and prosodic features. Such variability poses specific challenges for automatic speech recognition systems, as differences in pronunciation, intonation, and rhythm can significantly affect recognition accuracy. Therefore, the development of an STT system based on multi-dialectal speech data represents one of the pressing tasks in the field of computational linguistics.

MAIN PART. In the field of computational linguistics, the development and improvement of automatic speech processing technologies constitute one of the most significant and pressing scientific directions of modern research. Global scientific experience indicates that the idea of speech synthesis was first proposed in the 18th century by English scholars. In their early experiments, attempts were made to produce artificial sounds resembling human speech through the use of mechanical devices. However, these initial efforts were limited in terms of both accuracy and naturalness of the generated speech.

With the rapid advancement of computer technologies in subsequent periods, specialized systems for speech processing were developed. These systems, based on modeling the phonetic, acoustic, and linguistic properties of language, significantly expanded the possibilities for automatic speech analysis and synthesis. One notable milestone in this area dates back to the mid-20th century. In 1961, researchers at Bell Laboratories conducted a speech synthesis experiment using the IBM 704 computer, marking one of the first instances in which a machine generated sounds resembling human speech[1].

According to Dennis Klatt, early speech synthesis systems were primarily based on acoustic models of the human speech apparatus and operated using an approach known as formant synthesis[2]. Although such systems allowed for a certain degree of controllability, they were characterized by a relatively low level of naturalness in the produced speech.

[1] Sami L. Review of Speech Synthesis Technology // Espoo, 1999. P.104.
[2] Klatt, D. H. Review of text-to-speech conversion for English//*The Journal of the Acoustical Society of America*, 82(3), 1987. –Pp.737–793.

By the 1990s, speech synthesis had entered a new developmental stage with the emergence of concatenative synthesis technology. This approach involves generating new speech by concatenating pre-recorded speech segments. The unit selection synthesis method proposed by Hunt and Black significantly enhanced the naturalness of synthesized speech[3]. However, this method requires large-scale speech corpora, which led to the development of national speech databases in the United States, Japan, and European countries. For instance, the CMU Arctic Speech Database subsequently became a standard resource for numerous text-to-speech (TTS) systems.

In the 2000s, speech synthesis based on Hidden Markov Models (HMM-based speech synthesis) experienced substantial advancement. Tokuda et al. (2002) demonstrated that speech synthesis systems could be effectively developed through statistical modeling even with relatively limited data resources[4]. This stage served as an important scientific foundation, particularly for the development of TTS systems in low-resource languages.

Since 2016, the field of speech synthesis has undergone a paradigm shift. The WaveNet model, developed by Google DeepMind, introduced a neural network-based approach capable of directly generating raw audio waveforms[5]. By achieving a level of quality closely approximating natural human speech, WaveNet established a new benchmark for modern TTS systems.

Subsequently, the development of the Tacotron (2017) and Tacotron 2 models enabled the implementation of end-to-end speech synthesis. These models perform the entire process – from text input to audio signal generation – within a single unified neural network architecture.

At present, leading technology companies such as Amazon, Google, Microsoft, and Meta are actively conducting research on multilingual speech synthesis. Microsoft's Neural TTS technology has made it possible to model speech with

[3] Hunt, A. J., Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database // In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*. Vol. 1, 1996. –Pp. 373-376.
[4] Zen H., Tokuda K., Black A. W. Statistical parametric speech synthesis //speech communication. 2009. №. 11. – Pp.1039-1064.
[5] Aaron O., Dieleman S., Heiga Z., Karen S., Oriol V., Alex G., Kalchbrenner N., Andrew S., Koray K. WaveNet: a Generative Model for Raw audio. 2016.

human-like intonation and emotional expressiveness. Meanwhile, Google's multilingual TTS systems are capable of generating speech in multiple languages using a single model.

Global experience demonstrates that, in developing a speech synthesizer for the Uzbek language, the use of software packages based on a hybrid approach is particularly effective. First, Uzbek is classified as a low-resource language, and large-scale speech corpora as well as AI-based language models for this language remain insufficiently developed. Consequently, there arises a necessity to integrate multiple technological approaches within the speech synthesis process.

Second, achieving high accuracy in synthesis systems requires the application of classical rule-based frameworks, which play a significant role. However, such systems often depend on manual configuration and extensive linguistic expertise. In contrast, modern neural network-based approaches enable faster speech generation and significantly improve the naturalness of synthesized speech. Therefore, in order to ensure an optimal balance between accuracy and efficiency, the adoption of a hybrid model is considered a well-founded and effective solution.

In this study, Uzbek dialects are selected as the primary object of research, and the stages of constructing a large-scale audio database covering the dialects of 12 regions of the republic, as well as its linguistic annotation, are analyzed from a scientific perspective. This process serves to create a multi-dialectal linguistic resource for speech synthesis systems and to model the phonetic and dialectal features of the Uzbek language within a digital environment.

The Uzbek language is characterized by a rich diversity of dialects, which in dialectology are generally studied on the basis of three major dialect groups. From a regional perspective, the dialects of the 12 regions differ from one another in various linguistic features. The development of a speech synthesizer based on Uzbek dialects thus necessitates the creation of a comprehensive linguistic database grounded in these distinctions.

Within the framework of this research, the initial stage involves the development of an audio corpus, and the proposed system is structured into three main components[6]:

1. The process of audio corpus annotation.
2. The model training process.
3. The speech synthesis (inference) process.

These processes are closely interconnected, and the quality of data is monitored at each stage. In particular, the audio corpus annotation process was carried out using the Label Studio platform[7]. In this context, the audio data required for the system must first be accurately annotated. The primary objective is to prepare high-quality and consistently standardized audio–text transcript pairs.

During this process, audio recordings representing dialectal features from each region are collected and aligned with their corresponding textual scripts. These data are then uploaded to the Label Studio platform through a data pipeline. The annotator (a dialect specialist) listens to each audio recording and produces the corresponding textual transcription. In addition, relevant metadata is assigned, including the specific dialect affiliation of the respondent and the audio sample. The annotation is then reviewed for accuracy and consistency, and finally submitted through the “Submit” function.

[6] Abduraxmonova N., O‘rinboyeva R. O‘zbek tili nutqiy sintezatori uchun lingvistik ta’minot yaratish tamoyillari (monografiya). Globe Edit. 2026. –106 b.
[7] https://mohirdev.labelcraft.org

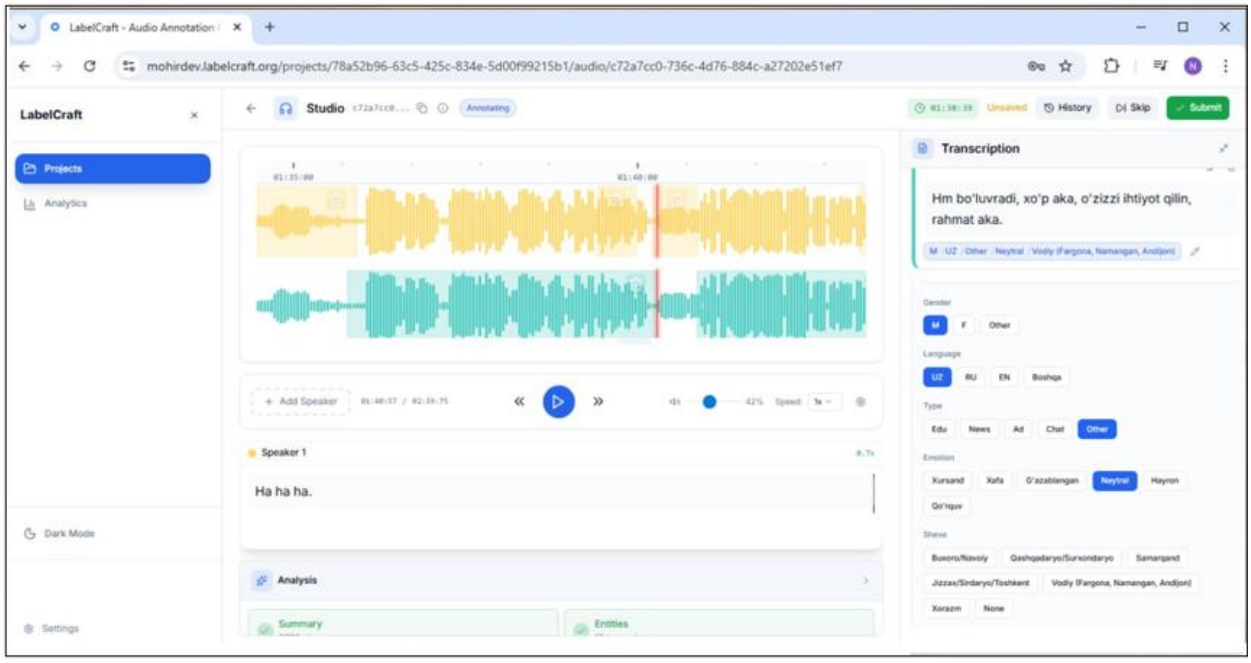


Figure 1. Linguistic annotation of audio in the Label Studio platform.

Completed work is then forwarded to the Reviewer (validator) for quality control. During the review stage, detected errors are corrected, or the annotation is returned to the annotator for revision. Upon completion of this process, the validated annotations become reliable material for developers. These data are subsequently used in the model training phase.

The next stage is the model training process, in which annotators and reviewers work collaboratively. Audio transcripts are prepared and thoroughly verified. The validated annotations are then used to train the model. The results are analyzed, and the model that demonstrates the best performance is selected.

One of the most critical stages in this process is audio data annotation. During annotation, speech recordings are segmented into semantically coherent units based on the speaker's linguistic characteristics. Particular attention is paid to speech signal quality, including minimal background noise, clarity of voice, and accuracy of pronunciation.

For each extracted segment from the full audio recording, relevant linguistic and extralinguistic parameters are assigned. These include the speaker's gender, language of communication, topic of speech, and emotional state inferred through

intonation (e.g., fear, joy, sadness, or neutrality), as well as the dialectal and regional affiliation of the speech. Segments with low intelligibility or unclear audio quality are excluded from the dataset using a dedicated “skip” function.

To improve the efficiency of the annotation process, a Speech-to-Text (STT) system based on artificial intelligence has been integrated into the platform, which generates preliminary transcriptions of segmented audio data. These automatically generated outputs are subsequently reviewed by annotators, edited, and aligned with the established linguistic annotation guidelines.

During the annotation stage, a set of specialized linguistic rules has been developed for each dialect, based on the principles of standard literary grammar. These rules ensure the consistent transcription of audio materials according to a unified standard while preserving the systematic representation of dialectal speech features. In particular, when lexical units belonging to other languages appear in the recordings, they are marked using a dedicated tagging system. For instance, Russian-language elements are annotated as [RU]…[/RU], while English-language elements are labeled as [ENG]…[/ENG].

However, lexical items that originate from other languages but are widely used in Uzbek, especially within dialectal speech and integrated into general communicative practice, are transcribed in accordance with their actual pronunciation in the respective dialect. This approach makes it possible to determine the functional usage level of such units within a specific dialectal context.

Similarly, abbreviated forms appearing in the audio are transcribed based on their phonetic realization rather than their written standard form. For example, the abbreviated form “SMS” is transcribed as “esemes” during the annotation process. Proper nouns, personal names, and geographical terms are written with initial capital letters in accordance with standard orthographic rules of the literary language. In addition, the representation of the phonemes x and h is carried out in accordance with the etymological form and the normative standards of the literary language.

Furthermore, a brief summary is provided for each annotated audio segment. This includes the identification of participants in the conversation (if applicable, their

names or professions) and a concise description of the main topic of the dialogue in 3-4 sentences.

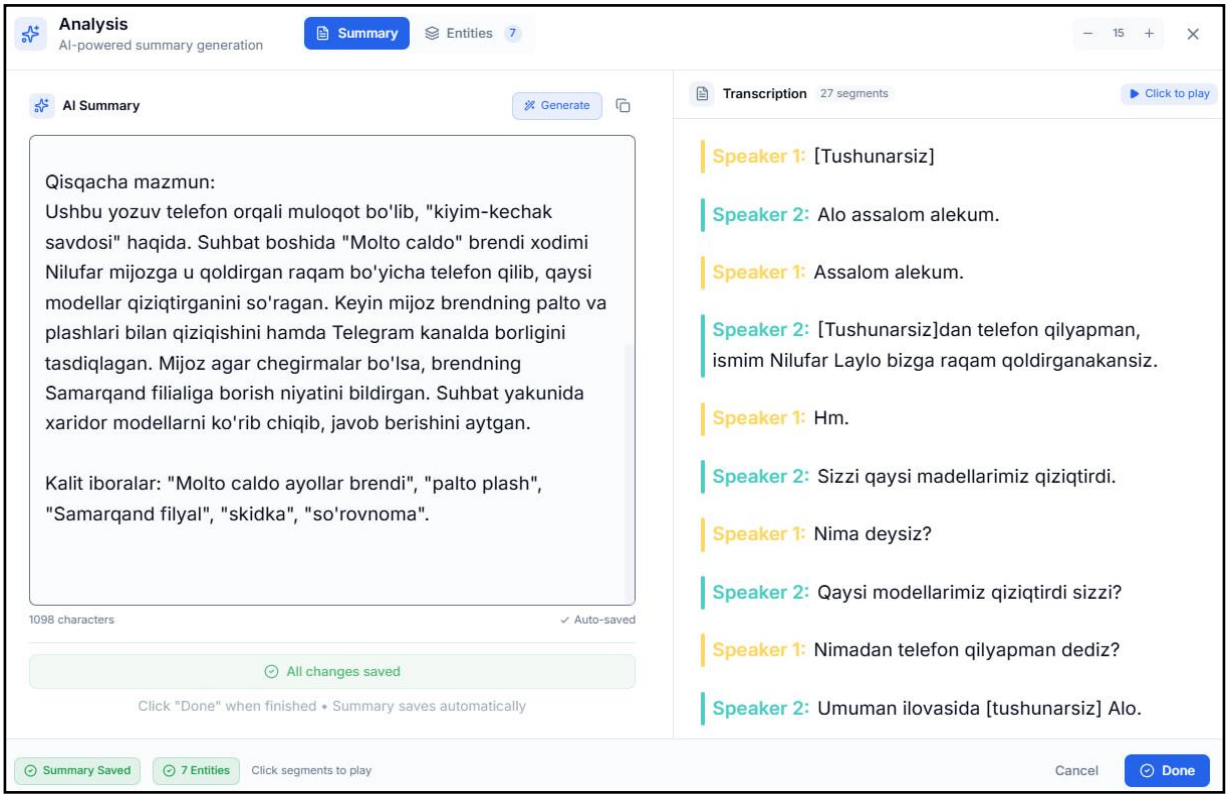


Figure 2. Generation of the summary section for audio during the annotation process.

In addition, keywords corresponding to the overall topic of the conversation are provided. The creation of a summary section facilitates the linkage of segmented units and enables their organization into coherent semantic groups.

CONCLUSION. The results of this study indicate that a well-structured linguistic annotation process plays a decisive role in the development of Uzbek dialect-based speech technologies. Segmenting audio data, defining its phonetic and extralinguistic features, and establishing unified annotation guidelines ensure the accuracy and stability of automatic speech processing systems. In particular, processing multi-dialect speech data through a standardized tagging system makes it possible to construct high-quality datasets for artificial intelligence models while preserving the dialectal diversity of the Uzbek language.

Furthermore, the integration of classical linguistic approaches with modern neural technologies within a hybrid framework provides an effective solution for low-

resource languages. As a result, the developed multi-dialect audio corpus contributes not only to improving speech synthesis and speech recognition systems but also to enriching the digital resources of the Uzbek language and advancing national language technologies.

References:

1. Abduraxmonova N.Z. Kompyuter lingvistikasi (darslik). –Toshkent: Nodirabegim, 2021. – 394 b.

2. Abduraxmonova N., O‘rinboyeva R. O‘zbek tili nutqiy sintezatori uchun lingvistik ta’minot yaratish tamoyillari (monografiya). Globe Edit. 2026. –106 b.

3. Aaron O., Dieleman S., Heiga Z., Karen S., Oriol V., Alex G., Kalchbrenner N., Andrew S., Koray K. WaveNet: a Generative Model for Raw audio. 2016.

4. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Upper Saddle River: Prentice Hall, 2009.

5. Klatt, D. H. Review of text-to-speech conversion for English//*The Journal of the Acoustical Society of America*, 82(3), 1987.

6. Hunt, A. J., Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database // In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*. Vol. 1, 1996.

7. Sami L. Review of Speech Synthesis Technology // Espoo, 1999. –104 p.

8. Zen H., Tokuda K., Black A. W. Statistical parametric speech synthesis //speech communication. 2009.

9. https://mohirdev.labelcraft.org

10. uzbekcorpus.uz