International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Neural network-based modeling for continuous speech recognition in the Uzbek Language

Narzillo Mamatov Solidjonovich [1], Nurbek Nuritdinov Davlataliyevich [1, *] and Muxiyatdinov Jamalatdin Kayratdin ulı [2]

[1] *Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan.*
[2] *(Basic Doctoral Program) at the National University of Uzbekistan.*

International Journal of Science and Research Archive, 2025, 16(01), 1539-1545

Publication history: Received on 09 June 2025; revised on 19 July 2025; accepted on 21 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2141

Abstract

This article presents a description of language models that take into account the unique features of the Uzbek language. Language modeling plays a significant role in improving the accuracy and performance of Automatic Speech Recognition (ASR) systems. Enhancing the conversion of speech to text can be achieved by correctly identifying syntactic and semantic structures in continuous speech. To achieve this goal, statistical and neural network-based language models, including deep learning architectures such as n-grams, Recurrent Neural Networks (RNNs), and transformer models, have been utilized.

Keywords: Automatic Speech Recognition (ASR); Language Model; Uzbek Speech; N-Gramm; Syntactic-Statistical Model; Neural Network Model; Uzbek Language Trigram Model; Hidden Markov Models (Hmms)

1. Introduction

The advancement of modern technologies has significantly facilitated human-computer interaction, with Automatic Speech Recognition (ASR) systems playing a crucial role [1]. These systems enable natural language communication with computers by automatically converting a user's speech into text [2]. The primary task of continuous speech recognition systems is to process human speech accurately and efficiently and represent it in a comprehensible textual format [3]. In this process, language modeling plays a vital role, allowing the system to understand the syntactic and semantic structures of speech and correctly interpret the context.

2. Literature review

We analyzed literature and articles on Automatic Speech Recognition (ASR), language model development, speech recognition, speech-to-text conversion, and machine translation methods.

2.1. Automatic speech recognition and uzbek language modeling

One of the key components of automatic speech recognition (ASR) systems is the language model. Many existing systems utilize statistical models based on n-gram words. These models estimate the probability of forming a sequence of *n* consecutive words in a text, making them effective for languages with rigid word order. However, this approach does not yield the expected results for languages like Uzbek, where word order is more flexible [4].

* Corresponding author: Nurbek Nuritdinov Davlataliyevich

Uzbek has several unique characteristics that reduce the efficiency of statistical models. As an agglutinative language with rich morphology, Uzbek's relatively free word order significantly increases the vocabulary size required for ASR systems and raises the ambiguity rate of n-gram language models. A language model represents the probability distribution of word sequences and is widely used in applications such as facilitating conversations, providing instant search results, improving translation quality, and assessing the meaning of social media messages. The better the language model, the better the outcomes, as evidenced by comparing older speech recognition and translation systems with modern ones [5].

Currently, several types of statistical language models exist to capture large-scale context or long-term relationships between words. One such type is *triggering patterns*, where the occurrence of a key word increases the likelihood of another word, referred to as the target, occurring.

A simplified form of trigger pairs is the *cache model*, which raises the probability of a word's occurrence based on how frequently it appears in the word history. This assumes that after using a particular word, the speaker is likely to reuse it, either because it is linked to a specific topic or due to the speaker's tendency to employ the word in their vocabulary.

A *long-distance trigram model* has also been proposed. This model predicts the probability of a word not only based on preceding words but also on those located further from the predicted word. It forms a set of word pairs that can be linked through certain "grammatical" separator words [6].

Another type of language model that allows for the modeling of long-distance relationships in a sentence is the *syntactic-statistical model*. To create such a model, statistical analysis of a text corpus is first conducted, and a list of n-gram words is compiled. Then, syntactic parsing is performed to identify grammatically related word pairs (syntactic groups) within the text. These syntactic groups are added to the list of n-gram words obtained through statistical analysis of the corpus. The process for creating a syntactic-statistical language model is illustrated in Figure 1.

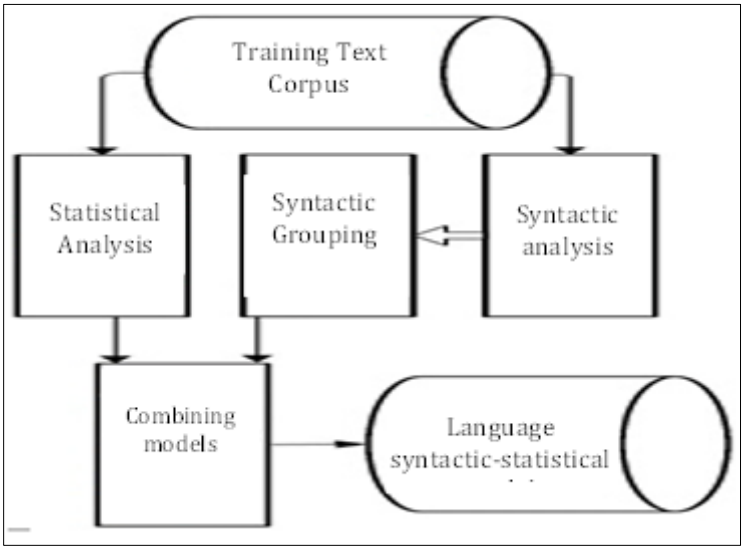

Figure 1 The process of creating a syntactic-statistical language model

Models based on particles are used for morphologically rich languages, such as inflected languages. Here, the word *w* is broken into specific parts (morphemes) through a function (L(w)). $U : w \to u^1, u^2, \ldots, u^{L(w)}$, $u^i \in \Psi$ is divided into, here Ψ Set of Word Parts. There are two methods for splitting words into morphemes: dictionary-based and algorithmic. The advantage of algorithmic methods is that they rely solely on text analysis and do not use any additional knowledge, making them applicable to any language. However, this results in words being split into pseudomorphemic units. The advantage of dictionary-based methods is that they allow for the correct division of words into morphemes, which can be used at the processing level in subsequent phrase recognition hypotheses.

Another model that can be applied to languages with rich morphology is the *Factorial Model for Language* (FML) for modeling the Arabic language. This model combines various characteristics (factors) of a word. $Y_i = \left(F_i^1, F_i^2, \ldots, F_i^k\right)$ It

is expressed as a *k* factor vector. A factor can be the shape of the word, part of speech, root, stem, and other morphological and grammatical features.

The language model can be built on artificial neural networks, specifically using a Recurrent Neural Network (RNN), which was used for the first time in this work. The advantage of this model is that the hidden layer preserves the context of all words prior to the current word being considered. The network consists of an input layer, a hidden layer (also known as the context layer or state), and an output layer. After training the neural network, the output layer provides the probability distribution of the previous and next words, taking into account the state of the hidden layer from the previous time step. The size of the hidden layer is typically chosen experimentally [7].

3. Results and discussion

This article presents the process of creating a factor and neural network model for the Uzbek language, as well as the use of these models for re-evaluating the best recognition hypotheses (N-best list) in automatic continuous speech recognition systems.

3.1. Uzbek Language Trigram Model

During the training process, a text corpus was collected from numerous websites, electronic journals, newspapers, and books with the goal of creating a model for the Uzbek language. Initial processing of the corpus was performed. As a result, a corpus with a total size of 250 million words was created, which was used for training the language's original (trigram) model, as well as factorial and neural network models.

To create the trigram language model, a specialized software tool was developed to automatically extract, process, and classify text from websites. Initial experiments on automatic continuous speech recognition in Uzbek showed that using a language model with a vocabulary of 120,000-word forms resulted in the fewest misrecognized words.

The evaluation of the created model for the Uzbek language was conducted on a text corpus that consisted of electronic informational resources not included in the training corpus. The ambiguity coefficient of the trigram language model was found to be 524 [8].

3.2. Uzbek Language Factor Model

The morphological analysis of the training text corpus was conducted using a specially developed tool. Word forms, lemmas, stems, parts of speech, and morphological tag features were applied, with all words replaced by their factors. For example, the word "uy" (house) would be transformed into "W-uylar, L-uylar, S-uylar, P-ot, M," where W stands for word form, L is the lemma, S is the stem, P is part of speech, and M represents all grammatical information related to the word.

Two-factor models were developed based on either the word form or one of the other factors mentioned above. If an n-gram does not exist in the training corpus or if its frequency is very low, it is replaced by the (n-1)-gram probability, which is multiplied by a back-off rate factor *y*. In n-gram models, back-off is done by discarding the longest word first, then the preceding word, and so on [9].

In factorial language models, two variants of the reverse process are possible, using two factors: first, the long word form and factor are discarded, followed by the nearest word form and factor (Figure 2.a), or the word forms can be sorted in distance order first, followed by the factors in the same order (Figure 2.b).

Based on this approach, the Uzbek language factor model was developed.

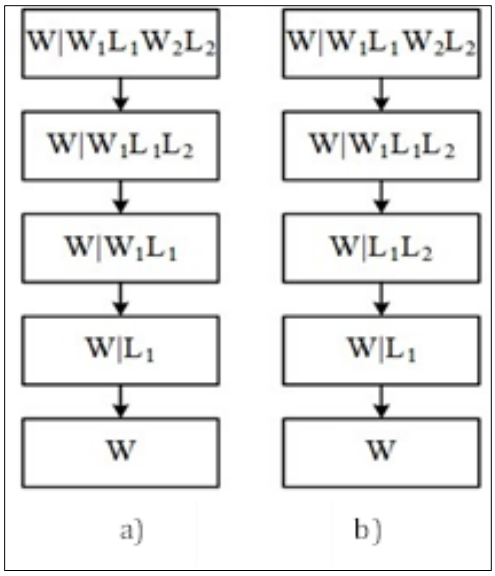

Figure 2 Back-off paths of the FNM characterized by word and lemma factors a) first back-off path; b) second back-off path

The ambiguity coefficients for the created FNM are provided in Table 1. Models with a back-off path of 1, regardless of the factors used, had lower ambiguity coefficients. Additionally, the model with the "word form" and "lemma" factors had the lowest ambiguity coefficient, with its back-off path equal to 1.

3.3. Uzbek Language Neural Network Model

Training for the RNN was carried out using the free software module based on the RNNLM Toolkit (Recurrent Neural Network Language Modeling Toolkit. To reduce the learning rate of the neural network, the output layer was separated into factors [10,11]. Classes are determined by word frequency. First, the probability distribution was computed based on the classes, and then the probability distribution for words belonging to the respective class was calculated. Models were created with 200, 500, and 800 hidden layer elements, and 200 and 800 classes.

Table 1 Uncertainty coefficient for FNM

FNM uncertainty coefficients		
Factors	**Uncertainty factor**	
	1 Path	**2 Path**
V.M.	457	561
W.L.	580	591
W.P.	682	736
W.S.	562	687

The values of the ambiguity coefficients for the developed models are presented in Table 2.

Table 2 Values of uncertainty coefficients for the models

Ambiguity Coefficients of the Uzbek Language Neural Network Models			
Number of Classes	**Number of Elements in the Hidden Layer**		
	200	500	800
200	960	991	8660
800	2080	860	880

Statistical Language Model for Continuous Speech Recognition in Uzbek. Hidden Markov Models (HMMs) with three states from left to right were used as the acoustic model, and these models were created using the HMM Toolkit (HMMT) based on a speech corpus derived from continuous speech of 60 speakers in Uzbek. The total size of the dataset is 60 GB, and the duration of the audio recordings exceeds 560 hours. The speech recognition system was tested using continuous spoken phrases taken from audiobooks. The sentences ranged from 3 to 30 words, and the recording time for each speaker was up to 30 minutes. In this case, the pure speech duration ranged from 10 to 25 minutes [12,13].

The total size of the test corpus is 1200 MB of audio data, and the automatic continuous speech recognition system for Uzbek was developed based on the transformer neural network model. The recognition system was evaluated based on the Word Error Rate (WER), which is a quality measure for speech recognition. During the decoding stage, a trigram language model was used, and the result was WER = 27.8% with hypothesis sizes of 10, 20, and 40. The hypotheses were then re-evaluated using the factor and language neural network models.

After re-evaluating the recognition results from the best recognition hypothesis list of the factor language model interpolated with a base language model with various interpolation coefficients [14], the results are provided in Table 3. Language model interpolation is the linear combination of word probabilities obtained from various models, taking into account the weight of each model. If only the factor language model is used, the interpolation coefficient is equal to 1; otherwise, it is less than 1.

Table 3 Recognition results after re-evaluating the list of best recognition hypotheses for the factorial model

The number of incorrectly recognized words (WER, %) after re-evaluating the best FNM recognition hypotheses list						
Language Model	**N=10**		**N=20**		**N=40**	
	Path 1	**Path 2**	**Path 1**	**Path 2**	**Path 1**	**Path 2**
WM+3-gr.	25.81	24.4	23.4	24.7	24.6	24.76
WL+3-gr.	25.8	25.1	25.5	25.3	25.52	25.38
WP+3-gr.	25.4	25.5	25.7	25.4	25.05	25.36
WS+3-gr.	25.2	26.2	25.87	25.63	25.87	26.12

Reevaluating the list of 20 hypothetical factor models, where word forms and morphological tags were used as factors, interpolated with a trigram model created with the backward path.

Experiments using the neural network models of the language, as with the experiments conducted with factor models, used the same list of good recognition hypotheses. Interpolation of the neural network and trigram models was also carried out, and the results of the experiments are shown in Table 2.4. Using the neural network model of the language allowed for a reduction in word recognition errors [15].

In the experiments, using a 200-class RNN showed better recognition results compared to using a 500-class RNN. The number of classes was set to 100, and an RNN with a hidden layer of 800 elements yielded the best result (WER = 23.7%) when interpolated with a trigram model with an interpolation coefficient of 0.6 based on the RNN-based language model."

Table 4 Experimental results of the interpolation of neural network and trigram models

After re-evaluating the list of different N-best hypotheses, the number of incorrectly recognized words (WER, %)				
Language model	**Interpolation coefficient, l**	**N=10**	**N=20**	**N=40**
An RNN with 200 elements in the hidden layer and a trigram NM are present	1.0	27.2	27.6	26.8
	0,6	26.3	25.6	25.9
	0,5	26.1	24.9	24.9
	0.4	25.0	24.7	24.6
An RNN with 400 elements in the hidden layer and a trigram NM are present	1.0	26.1	26.3	26.4
	0,6	25.8	25.3	24.1
	0,5	25.5	25.4	24.8
	0.4	25.3	24.9	24.2
An RNN with 800 elements in the hidden layer and a trigram NM are present	1.0	26.5	24.6	24.7
	0,6	23.7	24.7	22.9
	0,5	23.5	24.1	22.7
	0.4	23.8	24.2	23.6

4. Conclusion

Three different statistical models (trigram, factor, and neural network) of the Uzbek language have been extensively studied for continuous speech recognition systems in Uzbek. The advantage of the FNM over n-gram models is that the language model incorporates additional linguistic information, improving the quality of the speech recognition system for morphologically rich languages, including Uzbek. Models based on RNN are superior to other models as they retain arbitrary language context. Research on continuous speech recognition in Uzbek has shown that using factor and neural network models during the re-evaluation of the best recognition hypotheses helps reduce misrecognized words. Compared to results obtained using the language's trigram model, the relative reduction in misrecognized words was 10% when using the factor model and 18.5% when using the neural network model.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict-of-interest to be disclsoed.

References

[1] Abdinabi Mukhamadiyev, Mukhriddin Mukhiddinov, Ilyos Khujayarov, Mannon Ochilov and Jinsoo Cho. Development of Language Models for Continuous Uzbek Speech Recognition System. Sensors 2023, 23(3), 1145; https://doi.org/10.3390/s23031145.

[2] Mukhiddinov, M.; Akmuradov, B.; Djuraev, O. Robust Text Recognition for Uzbek Language in Natural Scene Images. In Proceedings of the 2019 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 4–6 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5. [Google Scholar].

[3] Musaev, M.; Khujayorov, I.; Ochilov, M. Automatic Recognition of Uzbek Speech Based on Integrated Neural Networks. In World Conference Intelligent System for Industrial Automation; Springer: Cham, Switzerland, 2021; pp. 215–223. [Google Scholar]

[4] Niyozmatova, N., Mamatov, Narzillo, Tulaganova, Sh., Samijonov, Abdurashid, and Samijonov, B. (2023). Methods for detecting speech activity in Uzbek speech in recognition systems. 050019. https://doi.org/10.1063/5.0145438.

[5] Mamatov, N.S., Niyozmatova, N.A., Yo'ldoshev, Y.S., Abdullaev, S.S., and Samijonov, A.N. (2023). Automatic speech recognition using an attention-based neural network. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (Eds.) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, Vol. 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_11.

[6] Mamatov, N.S., Niyozmatova, N.A., Samijonov, A.N., and Samijonov, B.N. (2022). "Development of language models for the Uzbek language," 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2022, pp. 1-4, https://doi.org/10.1109/ICISCT55600.2022.10146788.

[7] Mamatov, Narzillo, Niyozmatova, N., Abdullaev, Sh., and Samijonov, Abdurashid and Erejepov, K. (2021). Speech recognition based on transformer neural networks. 1-5. https://doi.org/10.1109/ICISCT52966.2021.9670093.

[8] Mamatov, N., Niyozmatova, N., and Samijonov, A. (2021). Software for preprocessing voice signals. International Journal of Applied Science and Engineering, 18, 2020163. https://doi.org/10.6703/IJASE.202103_18(1).006.

[9] Wiedecke, Bernd, Mamatov, Narzillo, Payazov, Mirabbos, and Samijonov, Abdurashid. (2019). Acoustic signal analysis and detection. International Journal of Innovative Technology and Exploring Engineering, 8, 2440-2442. https://doi.org/10.35940/ijitee.J9522.0881019.

[10] Narzillo, M., Abdurashid, S., Parakhat, N., and Nilufar, N. (2019). Automatic loudspeaker detection based on vector quantization method. International Journal of Innovative Technology and Exploring Engineering, 8(10), 2443–2445. https://doi.org/10.35940/ijitee.J9523.0881019.

[11] Mamatov, Narzillo and Niyozmatova, N. and Abdullaev, Sh and Samijonov, Abdurashid and Erejepov, K.. (2021). Speech Recognition Based on Transformer Neural Networks. 1-5. 10.1109/ICISCT52966.2021.9670093 Farncois Chollet "Deep learning with Python" s-386, 2016y.

[12] N. S. Mamatov, N. A. Niyozmatova, A. N. Samijonov and B. N. Samijonov, "Construction of Language Models for Uzbek Language," 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2022, pp. 1-4, doi: 10.1109/ICISCT55600.2022.10146788

[13] Niyozmatova, N. and Mamatov, Narzillo and Tulyaganova, Sh and Samijonov, Abdurashid and Samijonov, B. (2023). Methods for determining speech activity of uzbek speech in recognition systems. 050019. 10.1063/5.0145438.

[14] Mamatov, N.S., Niyozmatova, N.A., Yuldoshev, Y.S., Abdullaev, S.S., Samijonov, A.N. (2023). Automatic Speech Recognition on the Neutral Network Based on Attention Mechanism. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_11

[15] Jurafsky, D.; Martin, J.H. Speech and Language Processing. In An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Chapter-3, N-gram Language Models, 3rd ed.; Pearson: London, UK, 2014; pp. 29–55. [Google Scholar]