UDC 004.8
DOI: 10.53297/18293336-2025.1-28

DEVELOPMENT OF GRAPHEME-TO-PHONEME CONVERSION MODEL FOR ARMENIAN LANGUAGE

K.H. Nikoghosyan

National Polytechnic University of Armenia
"SYNOPSYS ARMENIA" CJSC

The paper presents the development of a grapheme-to-phoneme (G2P) conversion model for the Armenian language. The G2P process is a crucial component in the creation of Text-to-Speech (TTS) systems, directly impacting the quality of synthesized speech. Current approaches for Armenian G2P conversion demonstrate limited accuracy, as evidenced by high error rates of 96.60% WER and 36.15% PER in the existing tools like phonemizer. This research addresses these challenges by developing a comprehensive solution including a specialized dataset and neural network model. We begin by analyzing the specific phonological characteristics of Armenian, including context-dependent pronunciation rules and unique sound-symbol relationships that complicate automated transcription. To address the lack of publicly available resources, we have created a dataset containing 17,862 Armenian word-phoneme pairs by automatically collecting and processing data from Wiktionary using a multi-layered analysis system with robust quality control mechanisms. The analysis of this dataset revealed complex mapping patterns between Armenian graphemes and phonemes, with distribution characteristics following Zipf's law and a wide variety of contextual dependencies. Using this dataset, we developed a Conformer-CTC neural network model with approximately 12.3 million trainable parameters, featuring self-attention mechanisms and convolutional modules specifically designed to capture both local and global linguistic patterns. Evaluation shows that our model achieves a 16.13% Word Error Rate (WER) and a 17.36% Phoneme Error Rate (PER), representing an 80.47% and 18.79% improvement respectively over the existing solutions.

Keywords: Grapheme-to-Phoneme (G2P), Armenian language, neural networks, Text-to-Speech (TTS), Conformer-CTC, International Phonetic Alphabet (IPA).

Introduction. Text-to-Speech (TTS) systems have become essential tools for various applications, including accessibility services, virtual assistants, and educational platforms. A crucial component in TTS pipeline is the grapheme-to-phoneme (G2P) conversion process, which transforms the written text into its phonetic representation [1]. This transformation is particularly challenging for languages with complex orthographic systems or inconsistent spelling-pronunciation relationships.

The Armenian language presents unique challenges for G2P conversion due to its specific phonological features and the complexity of mappings between its writing system and pronunciation. While Armenian orthography is relatively consistent compared to languages like English, it still contains context-dependent rules and exceptions that complicate automated phonetic transcription [2].

The existing G2P solutions for Armenian, such as the phonemizer library [1], show significant limitations in accuracy and comprehensiveness. These limitations directly impact the quality of TTS systems, as incorrect phonetic transcriptions lead to unnatural or incomprehensible synthesized speech.

This paper addresses these challenges by developing a comprehensive G2P solution for Armenian, consisting of two main components: (1) a specialized dataset containing Armenian word-phoneme pairs, and (2) a neural network model based on the Conformer-CTC architecture. The dataset was created by automatically collecting and processing data from Wiktionary, focusing on words with IPA (International Phonetic Alphabet) transcriptions. The neural model was designed to effectively learn the complex mappings between Armenian graphemes and phonemes.

The remainder of this paper is organized as follows: Section 2 discusses the related work in G2P conversion, particularly for Armenian. Section 3 describes the methodology of dataset collection and the architecture of the proposed model. Section 4 presents the results of the evaluation and discusses their implications. Finally, Section 5 provides conclusions and directions for future work.

Related works. Grapheme-to-phoneme conversion has been an active area of research for decades, with approaches evolving from rule-based systems to statistical models and, more recently, neural networks.

Rule-based approaches rely on linguistic knowledge encoded in dictionaries and rule sets. While these methods can be effective for languages with regular orthography, they struggle with exceptions and require significant linguistic expertise to develop. For Armenian, rule-based approaches have been implemented in tools like espeak-ng, which is incorporated in the phonemizer library.

Statistical approaches utilize machine learning techniques to learn grapheme-phoneme mappings from data. These methods include Hidden Markov Models, Conditional Random Fields, and joint n-gram models. The Festival framework represents this category in the phonemizer library, requiring training data to function effectively.

Deep learning approaches have demonstrated superior performance in recent years. Sequence-to-sequence models with attention mechanisms have shown promising results across various languages. Transformer-based architectures have further improved performance by enabling better modeling of long-range dependencies in the input text.

For Armenian specifically, research on G2P conversion has been limited. The phonemizer library provides basic functionality but shows low accuracy, as demonstrated by evaluation metrics. The library includes three main engines: espeak-ng (rule-based), Festival (statistical), and Segments (table-based mapping). For Armenian, only the espeak-ng engine is available, resulting in limited performance.

Evaluation of phonemizer for Armenian shows a Word Error Rate (WER) of 96.60% and a Phoneme Error Rate (PER) of 36.15%, indicating significant room for improvement. This low accuracy directly impacts TTS quality, affecting rhythm, intonation, and pronunciation of synthesized speech.

The work presented in this paper builds upon these approaches while addressing their limitations for Armenian. By creating a comprehensive dataset and utilizing a state-of-the-art neural architecture, we aim to significantly improve G2P conversion performance for Armenian.

Materials and methods. Our approach to developing a G2P conversion system for Armenian consists of two main phases: (1) dataset collection and processing, and (2) neural model development and training.

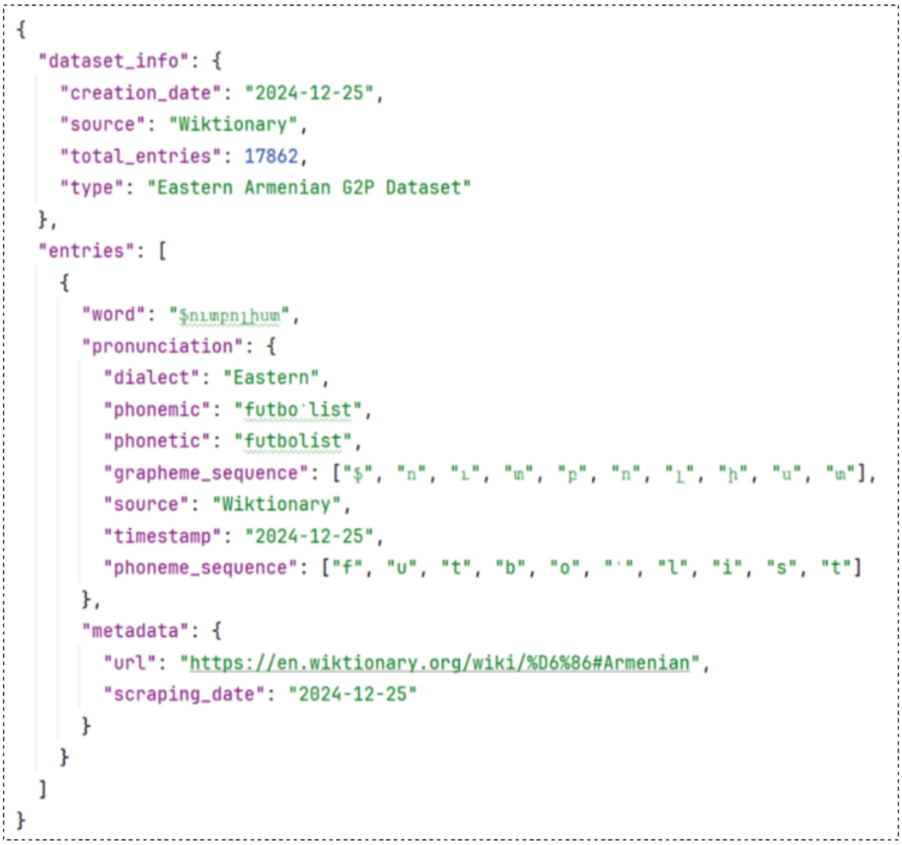
Dataset Collection and Processing. The lack of publicly available G2P datasets for Armenian necessitated creating our own dataset. We selected Wiktionary [3] as the primary data source, specifically targeting the "Armenian terms with IPA pronunciation" category [4]. This choice was motivated by several factors: Wiktionary's category system allows efficient identification of words with IPA transcriptions, and its page structure enables automated data collection.

For data collection, we developed a specialized Python-based system with a modular architecture. The collection module uses the BeautifulSoup library [5] for HTML structure analysis, efficiently extracting necessary information from word entries. The system implements a multi-layered analysis mechanism for each entry: first analyzing the general structure and isolating the Armenian section, then searching for and extracting phonological information, and finally identifying and processing the Eastern Armenian pronunciation variant.

The data cleaning and normalization process includes validating the phonetic transcription against IPA standards and processing special characters and diacritics that could affect training quality. The system also implements detailed logging, recording all operations including successful and failed queries, data processing stages, and identified issues.

Quality control is implemented through a multi-level verification mechanism. The first level performs automated checks to detect inconsistencies in data extracted from web pages. The second level verifies data completeness, confirming the presence of all mandatory fields. The third level implements content control, checking the semantic accuracy of phonetic transcriptions.

The collected data is stored in a specialized JSON format (Fig. 1), chosen for its hierarchical organization capabilities, human and machine readability, and widespread use in modern software systems. Each record includes the word and its phonetic transcription at the base level, grapheme and phoneme sequences in separate fields for easier processing, and metadata about the data source, collection date, and quality indicators.



```json
{
  "dataset_info": {
    "creation_date": "2024-12-25",
    "source": "Wiktionary",
    "total_entries": 17862,
    "type": "Eastern Armenian G2P Dataset"
  },
  "entries": [
    {
      "word": "ֆուտբոլիստ",
      "pronunciation": {
        "dialect": "Eastern",
        "phonemic": "futboˈlist",
        "phonetic": "futbolíst",
        "grapheme_sequence": ["ֆ", "ո", "ւ", "տ", "բ", "ո", "լ", "ի", "ս", "տ"],
        "source": "Wiktionary",
        "timestamp": "2024-12-25",
        "phoneme_sequence": ["f", "u", "t", "b", "o", "ˈ", "l", "i", "s", "t"]
      },
      "metadata": {
        "url": "https://en.wiktionary.org/wiki/%D6%86#Armenian",
        "scraping_date": "2024-12-25"
      }
    }
  ]
}
```

Fig. 1. An example of a dataset JSON structure

The final dataset includes 17,862 Armenian word-phoneme pairs, providing comprehensive coverage of the Armenian phonological system.

Model Architecture and Training. For Armenian G2P conversion, we developed a Conformer-CTC neural model [6]. As shown in Fig. 2, the model's core structure includes sequential processing layers from input feature formation to final phonetic predictions.

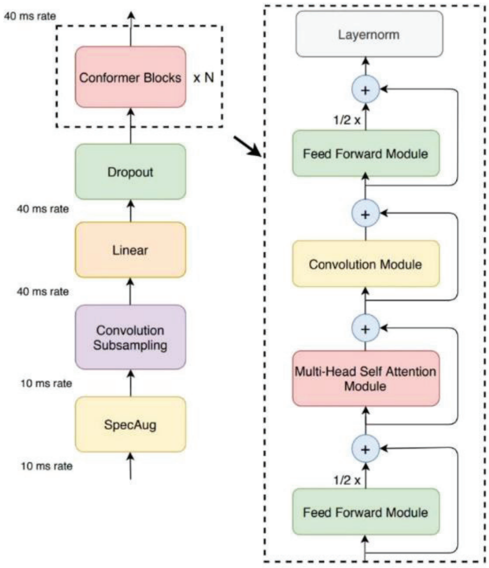


Fig. 2. The G2P-Conformer-CTC model architecture

The model's primary goal is to implement an effective mapping between the input grapheme sequence $x = (x_1, \dots, x_n)$ and the corresponding phoneme sequence $y = (y_1, \dots, y_m)$. This mapping is formulated as a probabilistic model:

$$p(y|x) = \sum_{\pi \in B^{-1}(y)} p(\pi|x), \qquad (1)$$

where $B^{-1}(y)$ represents all possible CTC paths corresponding to the phoneme sequence y.

The model's embedding layer transforms the input text into a 300-dimensional vector space. Each grapheme x_i is mapped to a vector e_i, encoding the grapheme's features and its possible phonological variants. The character set for Armenian includes the complete alphabet and punctuation marks.

The Conformer encoder structure is based on the relative positional self-attention mechanism [7], described by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + R\right)V, \qquad (2)$$

where R is the relative positional encoding matrix and d_k is the attention mechanism's dimensionality. The model uses 4 attention heads, allowing parallel processing of different linguistic features.

A key component is the CTC loss function [8], enabling training without direct grapheme-phoneme alignment annotation:

$$\mathcal{L}CTC = log\, p\ (y|x) = -\ log \sum \pi \in B^{-1}(y) \prod_{t=1}^{T} p(\pi_t|x)\ , \tag{3}$$

where T is the number of time steps, and $p(\pi_t|x)$ is the probability of a specific phoneme at a given time step.

The Conformer block's internal structure includes several main components: a convolutional module, self-attention module, and feed-forward network. The convolutional module's output is described by [9]:

$$H = GLU(W_1 * X + b_1) \otimes (W_2 * X + b_2), \tag{4}$$

where GLU is the output of the linear block, and $\otimes$ represents element-wise multiplication.

Normalization also plays an important role in model architecture. Each Conformer block applies batch normalization [10]:

$$BN(x) = \gamma \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}, \tag{5}$$

where μ_B and σ_B^2 are the batch mean and variance, and γ and β are learnable parameters.

For learning rate scheduling, we employed the Noam learning rate scheduler [10, 11]:

$$lr = d_{model}^{-0.5} \cdot min\ (step^{-0.5}, \qquad step \cdot warmup\_steps^{-1.5})\ , \tag{6}$$

where $d_m = 176$ is the model's main dimensionality, and $warmup\_steps = 10000$.

The overall model contains approximately 12.3 million trainable parameters, optimized using the AdamW optimizer [11] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$.

Training was performed on the created dataset, with a batch size of 32 and implementation of early stopping to prevent overfitting. The learning rate was managed by the Noam scheduler, allowing adaptive adjustment throughout the training process.

Results and Discussion. The training process was conducted on the dataset of 17,862 Armenian word-phoneme pairs, split into training, validation, and test subsets. Training dynamics (Fig. 3) shows consistent decrease in loss function for both training and validation sets. The model converges after approximately 1000 steps, with training loss reaching about 0.4 and validation loss stabilizing around 0.55.

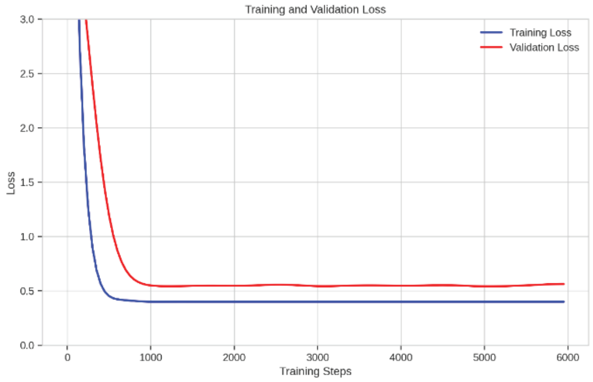


Fig. 3. Training and validation loss dynamics during training steps

Model effectiveness was evaluated using Word Error Rate (WER) and Phoneme Error Rate (PER). As shown in Fig. 4a, the model achieves a 16.13% WER on the validation set, correctly predicting 83.87% of words. This demonstrates high accuracy at the word level, especially considering the complexity of Armenian grapheme-phoneme correspondence.

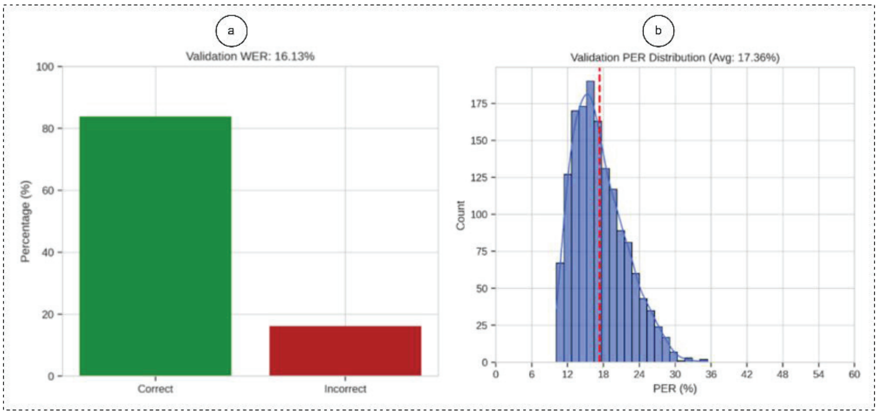


Fig. 4. WER distribution on the validation set(a). PER distribution on the validation set (b)

The PER distribution (Fig. 4b) shows a mean value of 17.36%. The distribution has a classic skewed nature, with the main mass concentrated in the 12-18% range, indicating that the model generally makes a small number of errors for each word. The distribution peak is around 15%, lower than the mean value, suggesting that the model performs better than average in many cases.

Compared to the existing solutions, the presented model shows significant improvement. The phonemizer library evaluation demonstrated a WER of 96.60% and PER of 36.15%. Our model reduces these error rates to 16.13% and 17.36%

respectively, representing a substantial improvement in accuracy. Spectral analysis shows that the model is particularly effective for:

- Simple grapheme-phoneme correspondences;
- Regular sound changes;
- Common word units;

However, challenges remain with rare words, complex phonological changes, and certain loanwords. These patterns are consistent with the behavior of similar models and indicate directions for future improvement.

Table presents a comparison of our model with the existing phonemizer tool, clearly demonstrating the superior performance of our approach across all evaluation metrics.

Table

Comparison of G2P conversion systems for Armenian

System	**WER (%)**	**PER (%)**	**Accuracy (%)**
Phonemizer (espeak-ng)	96.60	36.15	3.40
Our G2P-Conformer-CTC	16.13	17.36	83.87

The phoneme frequency analysis (Fig. 5) shows a distribution characteristic of natural languages, following the Zipf's law [12]:

$$P(r) = \frac{A}{r^{\alpha}}, \qquad (7)$$

where $P(r)$ is the frequency of the phoneme with rank r, r is the phoneme's ranking by frequency, α is a value close to 1, and A is a normalizing constant.

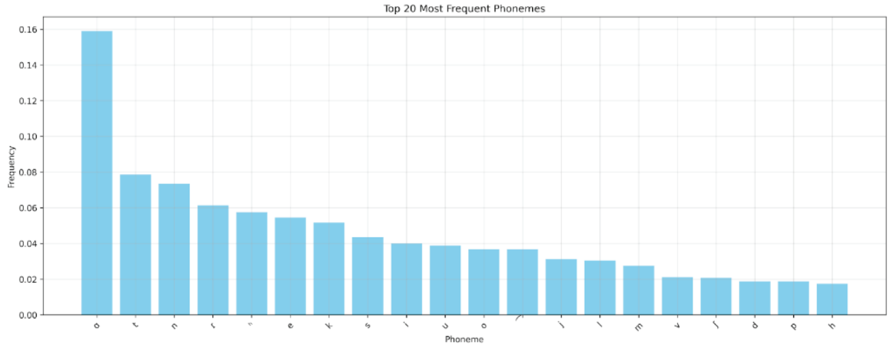


Fig. 5. Phoneme frequency distribution characteristic of Armenian

These findings highlight both the achievements and limitations of our approach. While the model demonstrates high accuracy for most words, there remain challenges with certain complex cases. Future improvements could focus on expanding the dataset, additional training for rare cases, and further optimization of model architecture.

Conclusion. This paper presented the development of a grapheme-to-phoneme conversion model for the Armenian language. We addressed the challenge of limited resources by creating a comprehensive dataset of 17,862 Armenian word-phoneme pairs and developing a Conformer-CTC neural network model tailored to the specific characteristics of Armenian phonology.

The results demonstrate significant improvement over the existing solutions, with our model achieving a Word Error Rate of 16.13% and a Phoneme Error Rate of 17.36%, compared to 96.60% and 36.15% respectively for the phonemizer library. This improvement directly impacts the quality of Text-to-Speech systems for Armenian, enabling more natural-sounding synthesized speech. Key contributions of this work include:

1. Creation of the first comprehensive, publicly available dataset for Armenian G2P conversion, providing a valuable resource for future research.
2. Development of a neural model architecture specifically designed for Armenian phonological characteristics.
3. Detailed analysis of Armenian grapheme-phoneme relationships, offering insights into the language's phonological structure.
4. Establishment of benchmark performance metrics for Armenian G2P conversion, enabling comparative evaluation of future systems.

The implications of this work extend beyond G2P conversion to the broader field the Armenian language processing. The created dataset and model can serve as foundations for developing various speech and language technologies, including speech recognition, language learning tools, and advanced TTS systems.

Future directions for this research include expanding the dataset to include more specialized terminology and neologisms, collecting a wider range of pronunciation variants, and enriching the dataset with stress and intonation information. The model architecture could be further refined to better handle complex cases and rare words.

This work represents a significant step toward advancing digital language technologies for Armenian, a language with limited computational resources. By providing open datasets and establishing performance benchmarks, we hope to stimulate further research and development in Armenian speech and language processing, ultimately contributing to the preservation and accessibility of the Armenian language in the digital age.

References

1. **Bernard M., and Titeux H.** Phonemizer: Text to Phones Transcription for Multiple Languages in Python // Journal of Open Source Software. - 2021. - Vol. 6. - P. 3958, doi: 10.21105/joss.03958.
2. The International Phonetic Association. https://www.internationalphoneticassociation.org.
3. Wiktionary - Multilingual online dictionary. https://en.wiktionary.org.
4. Armenian terms with IPA pronunciation - Category of Armenian words with IPA pronunciation. https://en.wiktionary.org//wiki/Category:Armenian_terms_with_IPA_pronunciation
5. BeautifulSoup - Python library for HTML parsing. https://www.crummy.com/software/BeautifulSoup/bs4/doc.
6. **Wu C., Sun H., Huang K., and Wu L.** MPSA-Conformer-CTC/Attention: A High-Accuracy, Low-Complexity End-to-End Approach for Tibetan Speech Recognition // Sensors. - 2024. - Vol. 24, no. 21. - P. 6824, doi: 10.3390/s24216824.
7. **Shaw P., Uszkoreit J., and Vaswani A.** Self-Attention with Relative Position Representations // ArXiv. – 2018, doi: 10.48550/arXiv.1803.02155.
8. **Graves A., Fernández S., Gomez F., and Schmidhuber J.** Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks // Proceedings of the 23rd International Conference on Machine Learning (ICML 2006) ._ 2006. - P. 369-376, doi: 10.1145/1143844.1143891.
9. Conformer: Convolution-augmented Transformer for Speech Recognition / **A. Gulati, J. Qin, C. Chiu, N. Parmar, et al** // ArXiv. – 2020, doi: 10.48550/arXiv.2005.08100.
10. **Ioffe S., and Szegedy C.** Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // ArXiv. - 2015. - doi: 10.48550/arXiv.1502.03167.
11. On the Variance of the Adaptive Learning Rate and Beyond / **L. Liu, H. Jiang, P. He, W. Chen, et al** // ArXiv. – 2019, doi: 10.48550/arXiv.1908.03265.
12. The Principle of Least Effort and Zipf Distribution / **Y. Zhu, B. Zhang, Q. Wang, et al** // Journal of Physics: Conference Series. - 2018. - Vol. 1113. - P. 012007, doi: 10.1088/1742-6596/1113/1/012007.

Received on 15.05.2025.
Accepted for publication on 10.07.2025.

ՀԱՅՈՑ ԼԵԶՎԻ ԳՐԱՆՇԱՆՆԵՐԸ ՀՆՉՅՈՒՆՆԵՐԻ ՎԵՐԱՓՈԽՄԱՆ ՄՈԴԵԼԻ ՄՇԱԿՈՒՄԸ

Կ.Հ. Նիկողոսյան

Ներկայացվում է հայոց լեզվի գրանշանները հնչյունների (ԳՀՎ) վերափոխման մոդելի մշակումը: ԳՀՎ գործընթացը կարևորագույն բաղադրիչ է տեքստից խոսքի վերափոխման (ՏԽՎ) համակարգերի ստեղծման գործում, ուղղակիորեն ազդելով սինթեզված խոսքի որակի վրա: Հայերենի ԳՀՎ փոխակերպման ներկա մոտեցումների ճշգրտությունը սահմանափակ է, ինչը հաստատվում է սխալների բարձր գործակիցներով՝ 96,60% ԲՍԳ և 36,15% ՀՍԳ, ինչպես օրինակ՝ phonemizer գործիքում: Այս հետազոտությամբ լուծվում են այդ մարտահրավերները: Մշակվել է համապարփակ լուծում, որը ներառում է մասնագիտացված տվյալների հավաքածու և նեյրոնային ցանցի մոդել: Նախ վերլուծում ենք հայերենի առանձնահատուկ հնչյունաբանական հատկանիշները, ներառյալ համատեքստից կախված արտասանության կանոնները, և յուրահատուկ հնչյուն-նշան կապերը, որոնք բարդացնում են ավտոմատացված տառադարձումը: Հանրային հասանելի ռեսուրսների բացակայության խնդիրը լուծելու համար ստեղծել ենք 17,862 հայերեն բառ-հնչյուն զույգ պարունակող տվյալների հավաքածու՝ Վիքիբառարանից տվյալները ավտոմատ հավաքագրելով և մշակելով բազմաշերտ վերլուծության համակարգի միջոցով՝ որակի հուսալի վերահսկման մեխանիզմներով: Այս տվյալների հավաքածուի վերլուծությունը բացահայտել է հայերեն գրանշանների և հնչյունների միջև բարդ համապատասխանեցման օրինաչափություններ, որոնց բաշխման բնութագրերը համապատասխանում են Ձիպֆի օրենքին և ունեն համատեքստային կախվածությունների լայն բազմազանություն: Օգտագործելով այս տվյալների հավաքածուն, մշակել ենք Conformer-CTC նեյրոնային ցանցի մոդել մոտավորապես 12,3 միլիոն ուսուցանվող պարամետրով, որն ունի ինքնաուշադրության մեխանիզմներ և փաթույթային մոդուլներ՝ հատուկ նախագծված տեղային և գլոբալ լեզվաբանական օրինաչափությունները բացահայտելու համար: Գնահատումը ցույց է տալիս, որ մեր մոդելն ապահովում է 16,13% բառային սխալների գործակից (ԲՍԳ) և 17,36% հնչյունային սխալների գործակից (ՀՍԳ), ինչը համապատասխանաբար 80,47% և 18,79% բարելավում է՝ գոյություն ունեցող լուծումների համեմատ:

Առանցքային բառեր. գրանշանները հնչյունների վերափոխող (ԳՀՎ), հայոց լեզու, նեյրոնային ցանցեր, տեքստը խոսքի վերափոխող (ՏԽՎ), Conformer-CTC, միջազգային հնչյունաբանական այբուբեն (ՄՀԱ):

РАЗРАБОТКА МОДЕЛИ ПРЕОБРАЗОВАНИЯ ГРАФЕМ В ФОНЕМЫ ДЛЯ АРМЯНСКОГО ЯЗЫКА

К.Г. Никогосян

Представлена разработка модели преобразования графем в фонемы (ГФП) для армянского языка. Процесс ГФП является ключевым компонентом при создании систем преобразования текста в речь (ПТР), напрямую влияя на качество синтезированной речи. Существующие подходы к преобразованию ГФП для армянского языка демонстрируют ограниченную точность, о чем свидетельствуют высокие показатели ошибок - 96,60% WER и 36,15% PER в существующих инструментах, таких как phonemizer. Данное исследование решает эти проблемы путем разработки комплексного решения, включающего специализированный набор данных и модель нейронной сети. Проведен анализ специфических фонологических характеристик армянского языка, включая контекстно зависимые правила произношения и уникальные звукосимвольные отношения, которые усложняют автоматизированную транскрипцию. Для решения проблемы отсутствия общедоступных ресурсов создан набор данных, содержащий 17,862 армянских пар слово-фонема, автоматически собирая и обрабатывая данные из Викисловаря с использованием многоуровневой системы анализа с надежными механизмами контроля качества. Анализ этого набора данных выявил сложные закономерности соответствия между армянскими графемами и фонемами с характеристиками распределения, соответствующими закону Ципфа, и широким разнообразием контекстуальных зависимостей. Используя этот набор данных, разработана нейросетевая модель Conformer-CTC с примерно 12,3 миллионами обучаемых параметров, включающая механизмы самовнимания и сверточные модули, специально разработанные для улавливания как локальных, так и глобальных лингвистических паттернов. Оценка показывает, что наша модель достигает 16,13% Word Error Rate (WER) и 17,36% Phoneme Error Rate (PER), что представляет собой улучшение на 80,47% и 18,79% соответственно по сравнению с существующими решениями.

Ключевые слова: графемно-фонемное преобразование (ГФП), армянский язык, нейронные сети, преобразование текста в речь (ПТР), Conformer-CTC, международный фонетический алфавит (МФА).