Proceedings of NPUA: Information technologies, Electronics, Radio engineering. 2025. № 2.
UDC 004.8
DOI: 10.53297/18293336-2025.2-38

ARMUS: A HIGH-QUALITY MULTISPEAKER ARMENIAN SPEECH CORPUS FOR SPEECH SYNTHESIS

K.H. Nikoghosyan

National Polytechnic University of Armenia
"SYNOPSYS ARMENIA" CJSC

The development of Text-to-Speech (TTS) systems requires high-quality speech datasets, which are particularly scarce for under-resourced languages like Armenian. This paper presents the development and implementation of an automated system for creating speech datasets from audiobooks and corresponding text files specifically designed for Armenian TTS applications. The system employs intelligent audio segmentation based on silence detection, text alignment mechanisms, and automated quality assessment protocols. Using this automated approach, a comprehensive Armenian speech dataset contain 14,182 audio segments was created with a total duration of 75,597.79 seconds (approximately 21 hours), sourced from professional audiobook recordings. The dataset includes recordings from two male speakers and covers 14,078 unique sentences containing 137,716 words with 30,466 unique vocabulary items. Audio files are standardized at 22,050 *Hz* sampling rate, 16-bit depth, and mono format to ensure consistency.

Quantitative analysis reveals that segment durations follow a natural distribution centered between 2-6 seconds, with an average duration of 5.33 seconds per segment. Phoneme distribution analysis demonstrates comprehensive coverage of the Armenian phonological system, following expected linguistic patterns. Quality assessment shows signal-to-noise ratios exceeding 35 *dB* across all segments, with 94.3% of randomly sampled segments meeting predefined quality criteria. The created dataset significantly exceeds existing Armenian speech resources in both volume and quality, providing a valuable foundation for Armenian TTS system development and other speech processing applications.

Keywords: Text-to-Speech (TTS), Armenian language, speech dataset, audio segmentation, silence detection, speech synthesis.

Introduction. Text-to-Speech synthesis has become an essential technology for various applications including accessibility services, virtual assistants, and language learning platforms. The development of high-quality TTS systems requires substantial amounts of paired speech and text data, which presents significant challenges for under-resourced languages. Armenian, despite its rich literary tradition and active speaker community, lacks sufficient publicly available speech datasets for advanced TTS system development.

The quality of TTS systems is directly dependent on the volume, diversity, and quality of training data [1]. While rule-based and statistical approaches can achieve reasonable results with limited data, modern neural TTS architectures require extensive paired audio-text datasets to achieve natural-sounding speech synthesis [2]. The scarcity of such resources for Armenian has hindered the development of high-quality TTS systems for this language.

The existing approaches to speech dataset creation often rely on manual segmentation and annotation processes, which are time-consuming, expensive, and prone to inconsistencies. Professional studio recordings, while offering superior audio quality, are typically limited in scope and vocabulary coverage. Conversely, crowd-sourced data collection can provide diversity but often suffers from quality control issues and inconsistent recording conditions.

This paper addresses these challenges by presenting an automated system specifically designed for creating high-quality speech datasets from existing audiobook resources. The system implements intelligent audio segmentation based on silence detection algorithms, automated text alignment, and comprehensive quality control mechanisms. By leveraging professionally recorded audiobooks, the approach combines the quality advantages of studio recordings with the efficiency of automated processing.

The primary contributions of this work include: (1) development of a modular automated system for speech dataset creation from audiobooks, (2) implementation of advanced silence detection and optimal segmentation algorithms, (3) creation of a comprehensive Armenian speech dataset with detailed quantitative analysis, and (4) establishment of quality benchmarks for Armenian speech data collection.

Related works. Speech dataset creation has been approached through various methodologies, each with distinct advantages and limitations. Professional studio recordings represent the gold standard for audio quality, exemplified by datasets such as LJSpeech for English [3] and CSS10 for multiple languages [4]. These datasets typically feature single speakers in controlled environments, resulting in consistent audio quality but limited speaker diversity.

Crowd-sourced approaches have gained popularity due to their scalability and cost-effectiveness. The Common Voice project [5] demonstrates this approach across multiple languages, collecting diverse speaker recordings through web-based platforms. However, such approaches face challenges in maintaining consistent audio quality and require extensive post-processing and validation.

Audiobook-based dataset creation represents a middle ground between studio quality and scalability. The LibriSpeech corpus [6] pioneered this approach for English, creating a large-scale dataset from audiobook recordings. Similar methodologies have been applied to other languages, including LibriTTS [7] which extends this approach with improved audio processing techniques.

For Armenian specifically, publicly available speech resources are extremely limited. The Mozilla Common Voice project includes some Armenian contributions, but the volume remains insufficient for training modern neural TTS systems [8]. Academic efforts have primarily focused on speech recognition rather than synthesis [9], leaving a significant gap in TTS-oriented resources.

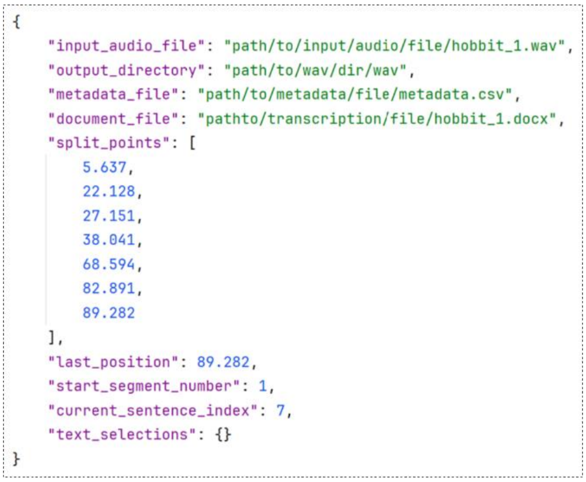
Automated segmentation techniques have evolved from simple silence-based approaches to sophisticated algorithms incorporating linguistic knowledge and machine learning. Voice Activity Detection (VAD) methods [10] form the foundation for most segmentation systems, while recent advances incorporate neural networks for improved accuracy in challenging acoustic conditions [11].

The work presented in this paper builds upon these established methodologies while addressing the specific challenges of Armenian speech processing including the language's unique phonological characteristics and the scarcity of the existing resources.

Materials and methods**.** The approach to creating a comprehensive Armenian speech dataset consists of two main phases: (1) development of an automated collection and processing system, and (2) systematic dataset creation and quality analysis.

Automated system development*.* The automated system employs modular architecture designed for efficient processing of audiobook materials. The system begins with configuration file creation or loading, guiding users through parameter selection and input file specification. Users must select input audio files, specify output directories for segmented audio files, define metadata file storage locations, and choose corresponding text files. The system also allows specification of initial segment numbering for organized file management.

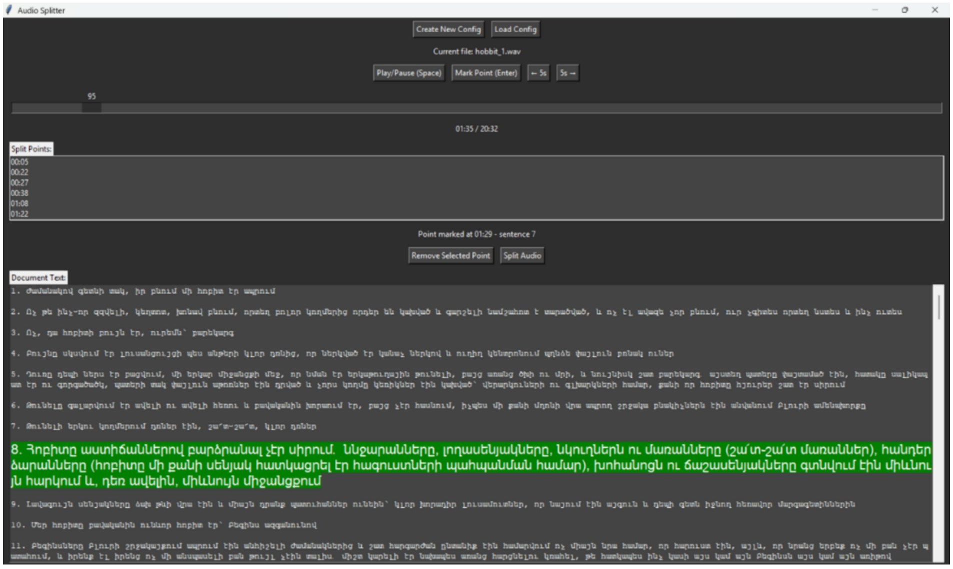
The configuration structure utilizes JSON format for hierarchical organization capabilities, human and machine readability, and widespread compatibility with modern software systems (Figure 1). Each configuration record includes audio and text file paths, output directory specifications, metadata storage locations, and processing parameters with quality control indicators.



```
{
    "input_audio_file": "path/to/input/audio/file/hobbit_1.wav",
    "output_directory": "path/to/wav/dir/wav",
    "metadata_file": "path/to/metadata/file/metadata.csv",
    "document_file": "pathto/transcription/file/hobbit_1.docx",
    "split_points": [
        5.637,
        22.128,
        27.151,
        38.041,
        68.594,
        82.891,
        89.282
    ],
    "last_position": 89.282,
    "start_segment_number": 1,
    "current_sentence_index": 7,
    "text_selections": {}
}
```

Fig. 1. A configuration file structure for the automated processing system

The core advantage of the proposed system is the automated audio segmentation based on silence detection. This approach enables natural-sounding segments while avoiding artificial speech interruptions. The graphical user interface allows visual monitoring of the segmentation process and necessary adjustments (Fig. 2).



Current file: hobbit_1.wav

01:35 / 20:32

00:05
00:22
00:27
00:38
01:08
01:22

Point marked at 01:29 - sentence 7

1. Ժամանակով գետնի տակ, իր բնում մի հոբիթ էր ապրում

2. Ոչ թե ինչ-որ զզվելի, կեղտոտ, խոնավ բնում, որտեղ բոլոր կողմերից որդեր են կախված և զարշելի նամշահոտ է տարածված, և ոչ էլ ավազե չոր բնում, ուր չգիտես որտեղ նստես և ինչ ուտես

3. Ոչ, դա հոբիթի բույն էր, ուրեմն՝ բարեկարգ

4. Բույնը սկսվում էր լուսանցույցի պես անթերի կլոր դռնից, որ ներկված էր կանաչ ներկով և ուղիղ կենտրոնում պղնձե փայլուն բռնակ ուներ

5. Դուռը դեպի ներս էր բացվում, մի երկար միջանցքի մեջ, որ նման էր երկաթուղային թունելի, բայց առանց ծխի ու մրի, և նույնիսկ շատ բարեկարգ. այստեղ պատերը փայտամած էին, հատակը սալիկապատ էր ու գորգածածկ, պատերի տակ փայլուն աթոռներ էին դրված և չորս կողմը կեռիկներ էին կախված՝ վերարկուների ու գլխարկների համար, քանի որ հոբիթը հյուրեր շատ էր սիրում

6. Թունելը գալարվում էր ավելի ու ավելի հեռու և բավականին խորանում էր, բայց չէր հասնում, ինչպես մի քանի մղոնի վրա ապրող շրջակա բնակիչներն էին անվանում Բլուրի ամենախորքը

7. Թունելի երկու կողմերում դռներ էին, շա՜տ-շա՜տ, կլոր դռներ

8. Հոբիտը աստիճաններով բարձրանալ չէր սիրում. Ննջարանները, լողասենյակները, նկուղներն ու մառանները (շա՜տ-շա՜տ մառաններ), հանդերձարանները (հոբիտը մի քանի սենյակ հատկացրել էր հագուստների պահպանման համար), խոհանոցն ու ճաշասենյակները գտնվում էին միևնույն հարկում և, դեռ ավելին, միևնույն միջանցքում

9. Լավագույն սենյակները ձախ թևի վրա էին և միայն դրանք պատուհաններ ունեին՝ կլոր խորադիր լուսամուտներ, որ նայում էին այգուն և դեպի գետն իջնող հեռավոր մարգագետիններին

10. Մեր հոբիթը բավականին ունևոր հոբիթ էր՝ Բեգինս ազգանունով

11. Բեգինսները Բլուրի շրջակայքում ապրում էին անհիշելի ժամանակներից և շատ հարգարժան ընտանիք էին համարվում ոչ միայն նրա համար, որ հարուստ էին, այլև, որ նրանցից երբեք ոչ մի բան չէր պատահում, և իրենք էլ իրենց ոչ մի անսպասելի բան թույլ չէին տալիս. Միշտ կարելի էր նախապես առանց հարցնելու կռահել, թե հատկապես ինչ կասի այս կամ այն Բեգինսն այս կամ այն առիթով

Fig. 2. The user interface of the automated data collection and processing system

The segmentation process implements a multi-stage algorithm. The first stage analyzes audio signal characteristics within ±1 second windows around each marked point. Silence detection employs a dynamic threshold method, calculating signal levels in decibels and comparing them against predefined thresholds.

The algorithm is based on multi-layer analysis. The first layer calculates signal levels for 50-millisecond windows. Signal levels are determined using the dBFS scale calculated for digital audio systems using the following formula:

$$\mathrm{dBFS} = 20\log_{10}\left(\frac{A}{A_{max}}\right), \tag{1}$$

where A is the absolute amplitude of the current sample, and A_{max} is the maximum possible amplitude in the digital system. For each window, dBFS is calculated as:

$$\mathrm{dBFS}(w) = 20\log_{10}\left(\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}}{A_{max}}\right), \tag{2}$$

where N is the number of samples in the window, and x_i represents individual samples. The silence detection criterion is defined as:

$$S(W) = \begin{cases} 1, dBFS(w) \ < \ -50 \\ \quad 0, otherwise \end{cases} \tag{3}$$

The second layer groups detected silent segments, identifying continuous silent regions. The third layer selects the central point of the longest silent region as the optimal segmentation point (OSP), determined by:

$$OSP = \frac{t_{start} + t_{end}}{2}, \tag{4}$$

where t_{start} and t_{end} represent the beginning and end of the longest silent segment.

The system supports automatic loading and analysis of DOCX format documents. For Armenian texts, special attention is given to proper handling of sentence segments separated by the ":" punctuation mark, which is characteristic of Armenian orthography.

A mechanism for merging short sentences has been developed based on a specially designed algorithm that analyzes the sentence length and word count. The minimum word count threshold is set based on empirical research showing that sentences containing fewer than 3 words often correspond to audio files shorter than 1 second. The algorithm also considers the minimum word length (4 characters) to exclude the influence of conjunctions and auxiliary words.

Audio file processing includes several important stages. Initial processing involves audio file loading and preliminary analysis. The system supports various formats (WAV, MP3) and automatically performs the necessary conversions to ensure uniform output formatting. All output audio segments are standardized to WAV format with 22,050Hz sampling rate, 16-bit depth, and mono configuration.

Dataset Creation and Quality Analysis. Using the developed system, a comprehensive Armenian speech dataset called ARMUS (Armenian Speech Dataset) was created for the TTS system training. The dataset source materials consisted of high-quality audiobook recordings from grqaser.org [12], specifically selecting books recorded by professional narrators with superior audio quality.

The system implements detailed logging mechanisms for subsequent analysis of segmentation effectiveness. The logging module preserves comprehensive information about each analysis, including marked timestamps, measured dBFS values, detected silent regions, and selected optimal points. This data are used for system optimization and parameter tuning.

Since each audio file requires a unique naming, a specialized segmentation mechanism is developed. The initial segment numbering selection mechanism enables flexible operation with large-scale data. The system automatically detects the existing segment numbers and suggests the next available number, allowing work with multiple files or different sections of the same file while maintaining numbering consistency.

Results and Discussion. The automated system successfully creates a comprehensive Armenian speech dataset with substantial improvements over the existing resources. The Table below presents the primary quantitative characteristics of the created dataset.

The dataset demonstrates substantial volume with approximately 21 hours of total audio, significantly exceeding the existing Armenian speech resources. The 14,182 audio files correspond to 14,078 unique sentences, with only 104 repeated sentences indicating dataset diversity and reducing overfitting probability during model training.

Table

Primary quantitative characteristics of the dataset

Metric	**Value**
Total duration (seconds)	75,597.79
Minimum audio file duration (seconds)	1.54
Maximum audio file duration (seconds)	16.47
Average duration (seconds)	5.33
Total quantity	14,182
Unique sentences	14,078
Total character count	869,097
Minimum characters per sample	12
Maximum characters per sample	347
Average characters per sample	61.28
Total word count	137,716
Unique words	30,466
Minimum words per sample	4
Maximum words per sample	56
Average words per sample	9.71
Number of speakers	2

Speaker Analysis. Special attention was given to speaker selection during data collection. The dataset includes recordings from two male speakers, with distribution shown in Figure 3. The "Gor" speaker contributes approximately 60% of the dataset (8,500 audio files), while "Narek" contributes 40% (5,400 audio files). This asymmetry reflects different source material volumes, with Gor's recordings sourced from "Ancient Greek Legends and Myths" and "The Count of Monte Cristo," while Narek's recordings come from "The Hobbit."

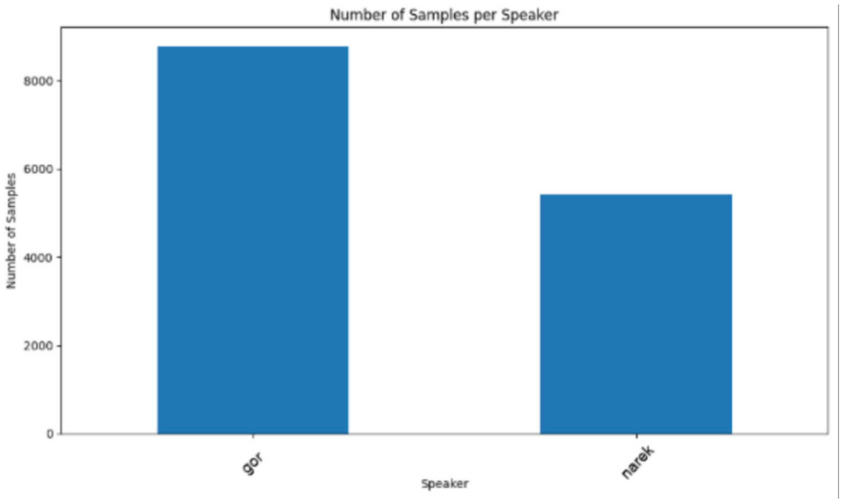


Fig. 3. Distribution of audio segments by speakers

Despite volume differences, both speakers provide sufficient data for effective multi-speaker model training. Speech quality analysis reveals consistent characteristics: Narek's average speech rate is 135 words per minute, while Gor's is 140 words per minute. This relatively small difference (approximately 3.7%) indicates speech tempo stability, positively affecting synthesized speech naturalness.

Duration analysis. The distribution of audio file durations presented in Fig. 4, shows that segments are primarily concentrated in the 2-6 second range, corresponding to natural single-sentence pronunciation duration. The shortest file duration is 0.48 seconds, while the longest is 16.47 seconds, with an average duration of 5.33 seconds.

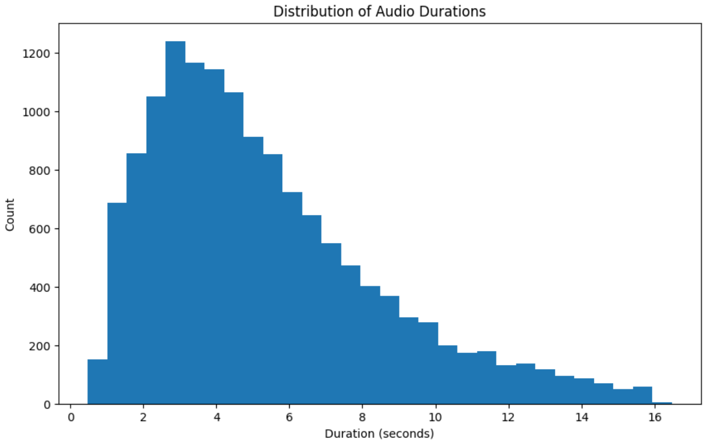


Fig. 4. Distribution of audio segment durations

This distribution is favorable for TTS system training, as most files are sufficiently long to enable intonational feature learning by neural models while remaining short enough to avoid training complications.

Linguistic analysis. The dataset's linguistic analysis demonstrates diversity and richness. The small number of repeated sentences (only 104) confirms dataset variety and reduces model overfitting probability during training. Special attention is given to balanced representation of Armenian phonemes in the dataset.

Phoneme distribution analysis, shown in Fig. 5, reveals correspondence with natural Armenian phonological distribution. The most frequently occurring phonemes are "ա", "ե", "ի", and "ն", which is characteristic of Armenian. This natural distribution is achieved through carefully selected literary texts ensuring comprehensive representation of the language's phonological system.

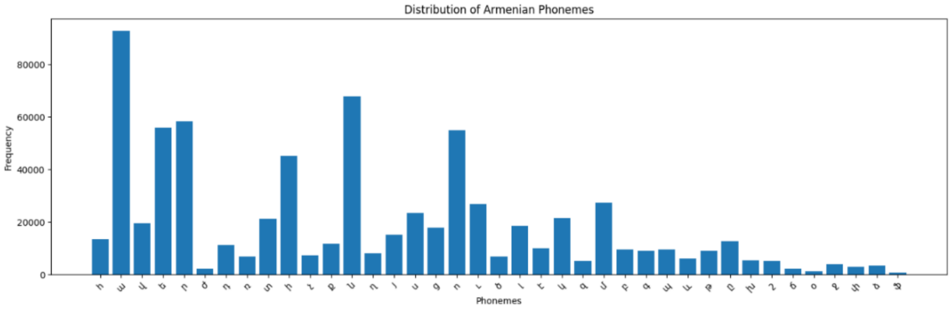


Fig. 5. Distribution of Armenian phonemes in the dataset

Quality Assessment. Audio quality is characterized by signal-to-noise ratio (SNR), calculated using the following formula:

$$SNR = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right), \qquad (5)$$

where P_{signal} is the useful signal power and P_{noise} is the noise power. SNR exceeds 35 decibels across all segments achieved through proper recording organization and post-processing in specialized sound-isolated studios using professional equipment.

Multi-stage quality verification was implemented during dataset creation. The first stage involved automated verification of all segments technical parameters through the automated system. The second stage implemented manual verification of 500 randomly selected segments, evaluating: (1) audio-text correspondence, (2) audio file boundary clarity, (3) recording quality and cleanliness, and (4) pronunciation clarity.

Verification results showed that 94.3% of segments fully met predefined criteria. The remaining 5.7% of identified deficiencies were primarily related to non-optimal segment boundary selection and were corrected through manual editing.

Comparative Analysis. To assess the applicability of the created dataset, comparative analysis was conducted with similar datasets for other languages. Research revealed that the dataset volume (21 hours) and quality characteristics are comparable to datasets used for training successful speech synthesis systems in other languages. For example, the LJSpeech dataset, widely used for English, contains approximately 24 hours of recordings.

The dataset's distinctive feature is genre diversity. Segments from "The Hobbit" contain rich dialogues and narrative sections, while "Ancient Greek Legends and Myths" are rich in descriptive texts. "The Count of Monte Cristo" provides narrative style diversity. This genre variety enables models to learn different speech stylistic characteristics.

Technical Standardization. To ensure the dataset technical quality and consistency, all recordings undergo strict standardization. Each audio file is preserved in WAV format with 22,050 *Hz* sampling rate and 16-bit depth. Original stereo recordings are converted to mono format, ensuring data uniformity and reducing memory consumption. Selected technical parameters balance audio quality with resource usage efficiency.

Applications and Limitations. The dataset's applications extend beyond speech synthesis due to high quality and meticulous annotation. It can be used for: (1) speech recognition system training, (2) phonetic model development, (3) prosody research, and (4) linguistic studies.

However, certain limitations must be noted. The small number of speakers (only two) may limit the synthesized voice diversity. Additionally, both speakers are male, limiting synthesis possibilities for female voice characteristics. These limitations can be addressed through future dataset expansion.

Conclusion. This paper presents the development and implementation of an automated system for creating high-quality speech datasets from audiobook resources, specifically applied to Armenian language materials. The automated approach combines professional audio quality with processing efficiency, resulting in a comprehensive dataset significantly exceeding the existing Armenian speech resources.

The created dataset, named ARMUS (Armenian Speech Dataset), contains 14,182 audio segments totaling approximately 21 hours, representing the first comprehensive and freely available Armenian speech corpus suitable for modern TTS system development. ARMUS is available for research purposes by contacting karen.nikoghosyan.98@gmail.com. Key achievements include: (1) successful implementation of intelligent audio segmentation based on silence detection, (2) creation of a balanced dataset with comprehensive phoneme coverage, (3) establishment of quality benchmarks for Armenian speech data collection, and (4) provision of a valuable resource for the research community.

The automated system's modular architecture enables adaptation to other audiobook sources and languages, potentially benefiting under-resourced language communities facing similar challenges. Technical standardization and comprehensive quality assessment ensure dataset reliability for various speech processing applications.

Our future work will focus on dataset expansion through additional speakers and genres, particularly including female speakers to enhance synthesis capabilities. Integration of stress and intonation information could further improve dataset utility for advanced TTS applications. The open availability of this dataset provides a foundation for collaborative development of Armenian speech technologies and contributes to digital language preservation efforts.

The significance of this work extends beyond technical contributions to broader goals of language technology democratization. By providing high-quality resources and establishing methodological frameworks, this research supports the development of speech technologies for Armenian while offering replicable approaches for other under-resourced languages.

References

1. **Taylor P.** Text-to-Speech Synthesis. - Cambridge University Press, 2009.
2. Tacotron: Towards End-to-End Speech Synthesis / **Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, et al** // ArXiv. - 2017, doi: 10.48550/arXiv.1703.10135.
3. **Ito K., and Johnson L.** The LJ Speech Dataset.- 2017. Available: https://keithito.com/LJ-Speech-Dataset/.
4. **Park K., and Mulc T.** CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages // ArXiv. - 2019, doi: 10.48550/arXiv.1903.11269.
5. Common Voice: A Massively-Multilingual Speech Corpus / **R. Ardila, M. Branson, K. Davis, et al** // ArXiv. - 2019, doi: 10.48550/arXiv.1912.06670.
6. **Panayotov V., Chen G., Povey D., and Khudanpur S.** Librispeech: An ASR corpus based on public domain audio books // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2015. - P. 5206-5210.
7. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech / **H. Zen, V. Dang, R. Clark, et al** // ArXiv. - 2019, doi: 10.48550/arXiv.1904.02882.

8. Mozilla Foundation. Common Voice Dataset. Available: https://commonvoice.mozilla.org/.
9. **Ghukasyan H., and Danielyan A.** Armenian Speech Recognition Using Deep Neural Networks // Proceedings of the International Conference on Computer Science and Information Technologies. - 2019.
10. **Sohn J., Kim N.S., and Sung W.** A statistical model-based voice activity detection // IEEE Signal Processing Letters. - 1999. - Vol. 6, no. 1. - P. 1-3.
11. **Zhang X.L., and Wang D.** A deep ensemble learning method for monaural speech separation // IEEE/ACM Transactions on Audio, Speech, and Language Processing. - 2016. - Vol. 24, no. 5. - P. 967-977.
12. Grqaser.org - Armenian Audiobook Platform. Available: https://grqaser.org/.

Received on 03.11.2025.
Accepted for publication on 29.01.2026.

ARMUS. ՀԱՅՈՑ ԼԵԶՎՈՎ ԲԱՐՁՐՈՐԱԿ ԲԱԶՄԱԽՈՍՆԱԿԱՅԻՆ ԽՈՍՔԱՅԻՆ ՏՎՅԱԼՆԵՐԻ ՇՏԵՄԱՐԱՆ՝ ԽՈՍՔԻ ՍԻՆԹԵԶԻ ՀԱՄԱՐ

Կ.Հ. Նիկողոսյան

Տեքստից խոսքի վերափոխման (ՏԽՎ) համակարգերի մշակումը պահանջում է բարձրորակ խոսքային տվյալների հավաքածուներ, որոնք հատկապես սակավ են նման սահմանափակ ռեսուրսներ ունեցող լեզուների, այդ թվում՝ նաև հայոց լեզվի համար։ Ներկայացվում են ավտոմատացված համակարգի մշակումը և ներդրումը՝ ձայնային գրքերից և դրանց համապատասխան տեքստային ֆայլերից խոսքային տվյալների հավաքածուներ ստեղծելու համար, որը հատուկ նախագծված է հայերեն ՏԽՎ կիրառությունների համար։ Համակարգը կիրառում է լռության հայտնաբերման վրա հիմնված խելացի ձայնային սեգմենտավորում, տեքստային համապատասխանեցման մեխանիզմներ և որակի ավտոմատ գնահատման արձանագրություններ։ Այս ավտոմատացված մոտեցման օգտագործմամբ ստեղծվել է համապարփակ հայերեն խոսքային տվյալների հավաքածու, որը պարունակում է 14,182 ձայնային սեգմենտ՝ 75,597.79 վայրկյան ընդհանուր տևողությամբ (մոտավորապես 21 ժամ), որոնք ստացվել են պրոֆեսիոնալ ձայնային գրքերի ձայնագրություններից։ Տվյալների հավաքածուն ներառում է արական սեռի երկու խոսնակների ձայնագրություններ և ընդգրկում է 14,078 եզակի նախադասություն, որոնք պարունակում են 137,716 բառ՝ 30,466 եզակի բառային միավորներով։ Ձայնային ֆայլերը ստանդարտացված են 22,050 *Hz* նմուշառման հաճախականությամբ, 16-բիթ խորությամբ և մոնո ձևաչափով։

Քանակական վերլուծությունը բացահայտում է, որ սեգմենտների տևողությունները հետևում են բնական բաշխվածությանը, որը կենտրոնացած է 2…6 վայրկյան միջակայքում, միջինը 5,33 վայրկյան մեկ սեգմենտի համար։ Հնչյունների բաշխվածության վերլուծությունը ցույց է տալիս հայոց հնչյունաբանական համակարգի համապարփակ ծածկույթը՝ հետևելով սպասելի լեզվաբանական օրինաչափություններին։ Որակի գնահատումը ցույց է տալիս՝ ազդանշան-աղմուկ հարաբերակցությունները գերազանցում են 35 *dB* բոլոր սեգմենտներում, իսկ պատահականորեն ընտրված սեգմենտների 94,3%-ը համապատասխանում է նախապես սահմանված որակի չափանիշներին։ Ստեղծված տվյալների հավաքածուն զգալիորեն գերազանցում է առկա հայերեն խոսքային ռեսուրսները ինչպես ծավալով, այնպես էլ որակով՝

տրամադրելով արժեքավոր հիմք հայերեն ՏԽՎ համակարգերի մշակման և խոսքի մշակման այլ կիրառությունների համար:

Առանցքային բառեր. տեքստից խոսքի վերափոխում (ՏԽՎ), հայոց լեզու, խոսքային տվյալների հավաքածու, ձայնային սեգմենտավորում, լռության հայտնաբերում, խոսքի սինթեզ:

АРМУС: ВЫСОКОКАЧЕСТВЕННЫЙ МНОГОДИКТОРСКИЙ КОРПУС АРМЯНСКОЙ РЕЧИ ДЛЯ СИНТЕЗА РЕЧИ

К.Г. Никогосян

Разработка систем преобразования текста в речь (ПТР) требует высококачественных наборов речевых данных, которые особенно редки для языков с ограниченными ресурсами, таких как армянский. В данной статье представлены разработка и внедрение автоматизированной системы для создания наборов речевых данных из аудиокниг и соответствующих текстовых файлов, специально разработанной для армянских ПТР-приложений. Система использует интеллектуальную сегментацию аудио на основе обнаружения тишины, механизмы выравнивания текста и протоколы автоматической оценки качества. Используя данный автоматизированный подход, создан всеобъемлющий набор данных армянской речи, содержащий 14,182 аудиосегмента общей продолжительностью 75,597.79 *с* (приблизительно 21 час), полученных из профессиональных записей аудиокниг. Набор данных включает записи двух дикторов мужского пола и охватывает 14,078 уникальных предложений, содержащих 137,716 слов с 30,466 уникальными словарными единицами. Аудиофайлы стандартизированы с частотой дискретизации 22,050 *Гц*, 16-битной глубиной и моноформатом для обеспечения согласованности.

Количественный анализ показывает, что продолжительности сегментов следуют естественному распределению с центром в диапазоне 2…6 *с*, со средней продолжительностью 5,33 *с* на сегмент. Анализ распределения фонем демонстрирует всестороннее покрытие армянской фонологической системы, следуя ожидаемым лингвистическим закономерностям. Оценка качества показывает отношения сигнал/шум, превышающие 35 *дБ* во всех сегментах, при этом 94,3% случайно отобранных сегментов соответствуют предопределенным критериям качества. Созданный набор данных значительно превосходит существующие армянские речевые ресурсы как по объему, так и по качеству, обеспечивая ценную основу для разработки армянских ПТР-систем и других приложений обработки речи.

Ключевые слова: преобразование текста в речь (ПТР), армянский язык, набор речевых данных, сегментация аудио, обнаружение тишины, синтез речи.