Exploring Armenian Speech Recognition: A Comparative Analysis of ASR Models - Assessing DeepSpeech, Nvidia NeMo QuartzNet, and Citrinet on Varied Armenian Speech Corpora

Varuzhan Baghdasaryan
National Polytechnic University of Armenia
Yerevan, Armenia
e-mail: varuzh2014@gmail.com

***Abstract*— In the dynamic landscape of speech recognition technology, the pursuit of flawless and precise recognition across diverse languages remains a steadfast goal. This study focuses on Armenian Speech Recognition, driven by the need to address robust ASR models for under-resourced languages. The research evaluates three prominent ASR models - DeepSpeech, Nvidia NeMo QuartzNet, and Citrinet - aiming to enhance Armenian acoustic and language models for higher accuracy.**

Armenian, an independent branch of the Indo-European family, holds a profound linguistic heritage, yet lacks tailored ASR solutions, posing significant challenges for accurate and contextually relevant speech recognition. The study aims to bridge this gap by thoroughly evaluating the capabilities and architectural design of the three ASR models.

DeepSpeech, an open-source deep recurrent neural network (RNN)-based ASR system by Mozilla, represents versatility and accuracy. On the other hand, Nvidia NeMo offers QuartzNet and Citrinet, lightweight and efficient ASR models optimized for real-time applications and edge devices.

The evaluation uses three datasets: ArmSpeech, Speech corpus of Armenian question-answer dialogues, and Google's FLEURS. These high-quality Armenian speech corpora form the basis for training, validating, and testing the ASR models, enabling a fair comparison of their capabilities and performance.

The research findings promise to unlock the true potential of Armenian Speech Recognition, enriching human-machine interaction and inspiring future ASR innovations.

***Keywords*— Armenian ASR, speech recognition system, speech-to-text, acoustic model, language model, DeepSpeech, Nvidia NeMo, QuartzNet, Citrinet, Armenian speech corpora.**

I. Introduction

ASR, a pivotal component of natural language processing, has transformed communication with machines. To optimize NLP engines, multi-speaker corpora training is recommended, incorporating diverse linguistic features for better efficiency across genders, ages, and accents.

This research focuses on developing and comparing two critical ASR models: the acoustic model [1], linking speech audio signals and phonemes, and the language model [2], understanding language words and their arrangement for exceptional prediction accuracy. Fine-tuning techniques will enhance their performance further.

The language model training utilizes the powerful KenLM library [3] for constructing n-gram language models [4].

Three formidable models, Mozilla's DeepSpeech [5, 6], Nvidia NeMo QuartzNet [7], and Citrinet [8], are carefully selected for acoustic model creation and evaluation.

DeepSpeech excels in accuracy and adaptability, ideal for multilingual and multi-dialectal tasks.

Nvidia NeMo QuartzNet is lightweight yet powerful, perfect for real-time and edge device applications.

Nvidia NeMo's Citrinet combines CNNs and transformer-based architectures for superior performance in context-dependent scenarios.

Comprehensive evaluation of ArmSpeech [9, 10], Speech corpus of Armenian question-answer dialogues [11], and Google's FLEURS [12] corpora will uncover valuable insights into their capabilities, strengths, and applications, driving advancements in Armenian Speech Recognition.

II. Datasets

The evaluation of ASR models will be conducted using the comprehensive ArmSpeech dataset, along with the Speech corpus of Armenian question-answer dialogues and Google's FLEURS corpora.

The ArmSpeech dataset, meticulously crafted, consists of 18 unique sections with a total duration of 15.7 hours [9, 10]. The first 13 sections encompass audio clips sourced from public-domain audiobooks of fiction, making it a strategic choice as it includes both common and less common words, enriching the dataset for training purposes [9, 10]. The 14th section includes speeches covering a diverse array of real-life

situations, such as movies, sports, restaurants, numbers, days of the week, and months [9, 10]. The remaining four sections were collected from volunteers and automatically annotated using a Python application, ensuring data integrity and consistency [9, 10]. To maintain optimal sound quality and facilitate audio editing, the audio clips were thoughtfully released in a high-quality format [9, 10]. These mono-channel 16-bit files boast a 16000 Hz sampling rate and 256 kbps bit rate, employing the WAV format with lossless compression [9, 10]. The diversity of the ArmSpeech dataset extends to its voice contributors, with 13 male (63.69 %) and 4 female (36.31 %) voices included, further enriching the dataset [9, 10].

The Armenian question-answer dialogues represent a valuable corpus of elicited controlled speech, meticulously curated to investigate intonation patterns within two Armenian dialects: Western Armenian (WA) and Eastern Armenian (EA) [11]. Designed with care, the corpus comprises dialogues interspersed with intermittent fillers, strategically crafted to act as stimuli and evoke desired variations in intonation. This resource proves highly beneficial for researchers delving into the realms of intonation prosody, forced alignment, and Automatic Speech Recognition (ASR) studies, offering opportunities to investigate the nuances of intonation in questions and answers within the context of under-resourced languages. The dataset is substantial, comprising a total of 8,852 dialogues, encompassing 23,711 individual sound files, amounting to a data size of 2.7GB and an audio duration of approximately 8.5 hours [11]. Each utterance is meticulously associated with a sound file, a Praat TextGrid file enriched with comprehensive linguistic annotations, and a text file containing orthographic forms to facilitate ASR-related tasks [11]. Moreover, pronunciation dictionaries are thoughtfully included, extending support to researchers engaged in ASR or forced alignment endeavors. The openly accessible nature of this dataset underscores its importance in fostering advancements in Armenian Speech Recognition and linguistics, promoting collaboration and offering deeper insights into intonation patterns within under-resourced languages, thereby paving the way for novel discoveries and breakthroughs in the field.

FLEURS, developed by Google, is an n-way parallel speech dataset encompassing 102 languages [12]. It is skillfully built upon the machine translation FLoRes-101 benchmark, offering approximately 12 hours of speech supervision per language [12]. This versatile resource caters to a wide range of speech-related tasks, including Automatic Speech Recognition (ASR), Speech Language Identification (Speech LangID), Translation, and Retrieval, making it a valuable addition to the evaluation process.

To ensure seamless integration of The Armenian question-answer dialogues and FLEURS corpora with DeepSpeech, QuartzNet, and Citrinet, Python scripts were employed to convert their formats accordingly. Specifically, adjustments were made to align the datasets with the target models' requirements. For The Armenian question-answer dialogues, only Eastern Armenian samples with corresponding labels were retained, ensuring perfect alignment with the target model's needs. In the case of FLEURS, a thorough manual examination was conducted to identify and address any defects or errors present in the corpus. This meticulous process involved manually removing audio clips that were either excessively long or short or consisted of damaged audio. Moreover, audio clips with inaccurately labeled transcriptions were corrected manually.

Furthermore, all audio clips were converted to mono-channel 16-bit files with a 16000 Hz sampling rate and 256 kbps bit rate, using the WAV format with lossless compression. After applying these necessary modifications, The Armenian question-answer dialogues encompassed 12,422 samples, spanning a total duration of 4.48 hours. On the other hand, the FLEURS dataset contained 4,380 samples, with a combined duration of 14.58 hours. To enable comprehensive evaluation and ensure a fair distribution of data, the final combined dataset, comprising these three datasets, was randomly divided into training, development, and testing sets, adhering to a ratio of 60-20-20%.

This meticulous dataset preparation process ensures the harmonious integration of the diverse datasets, allowing for meaningful and accurate evaluations of the ASR models' capabilities and performance. By fine-tuning the datasets to meet the specific requirements of DeepSpeech, QuartzNet, and Citrinet, this research paves the way for reliable and insightful investigations in the realm of Armenian Speech Recognition, ultimately contributing to advancements in this field and fostering further progress in linguistics and ASR research.

III. METHODS

At the heart of any Automatic Speech Recognition (ASR) system lies the harmonious collaboration between the acoustic and language models, two interdependent pillars that breathe life into the technology. The acoustic model serves as the "ears" of the system, unraveling the rich tapestry of speech audio and deciphering the phonetic or subword units that compose it. Drawing inspiration from the complex workings of human auditory perception, acoustic models employ sophisticated deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to capture the intricate temporal patterns that define spoken language.

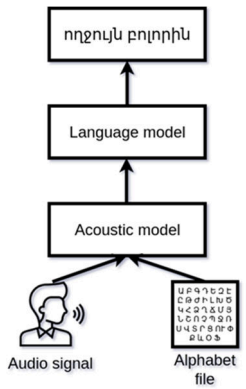

Figure 1. The fundamental concept of the modern ASR systems

Simultaneously, the language model acts as the "linguistic compass," skillfully navigating the vast terrain of human language and predicting the most probable word sequences based on their contextual relationships. This linguistic journey is facilitated by advanced modeling approaches, including n-gram techniques, recurrent neural networks (RNNs), and transformer-based architectures, which are capable of encapsulating the nuances of language in its purest form.

Mozilla's implementation of Baidu's DeepSpeech stands as a cutting-edge, open-source ASR system that has garnered widespread acclaim for its remarkable accuracy and versatility. DeepSpeech harnesses the power of deep learning, specifically employing a recurrent neural network (RNN), which enables it to capture context information from past and future time steps, making it highly effective for sequential data like speech [5, 6].

One of the most notable aspects of Mozilla's DeepSpeech is its flexibility, allowing developers to fine-tune pre-trained models on specific datasets. This adaptability renders it suitable for various domains and accents, enabling the recognition of speech from diverse sources and speakers. The model has been trained on an extensive range of multilingual and multi-dialectal speech data, contributing to its robustness and its ability to generalize across different languages.

Leveraging powerful libraries and frameworks like TensorFlow and PyTorch, Mozilla's DeepSpeech efficiently implements deep learning models. It employs various techniques, including transfer learning and data augmentation, to enhance the ASR system's performance and improve its ability to adapt to new environments.

The implementation of Mozilla's DeepSpeech incorporates Baidu's Deep Speech ASR and employs TensorFlow. The model consists of a deep recurrent neural network (RNN) composed of 5 hidden layers, adeptly performing supervised learning tasks [5, 6]. The output layer of the neural network is a fully connected layer that employs the softmax activation function [5, 6]. This function is used for normalization, enabling the layer to generate probabilities for each character in the language's alphabet.

This combination of hidden layers enables the neural network to process and extract meaningful information from the audio data effectively, contributing to the acoustic model's ability to decode speech signals accurately.

The architecture of Mozilla's implementation of Baidu's Deep Speech is illustrated in Figure 2.

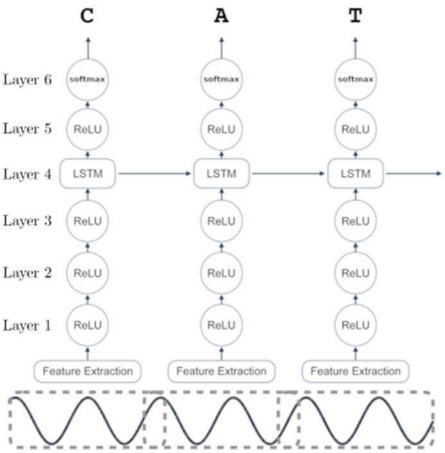

Figure 2. The architecture of Mozilla's Deep Speech [5, 6]

Nvidia NeMo's QuartzNet represents another remarkable ASR model, distinguished by its lightweight and efficient design, making it particularly well-suited for real-time applications and edge devices. Unlike conventional ASR models relying on recurrent neural networks (RNNs), QuartzNet adopts fully convolutional neural networks (CNNs) and 1D temporal convolutions [7]. This innovative approach enables the model to process input speech in parallel, resulting in a significant reduction in computational complexity while maintaining high efficiency and accuracy.

To further enhance computational efficiency, QuartzNet incorporates cutting-edge techniques such as depthwise separable convolutions and bottleneck modules [7]. These advancements effectively reduce the number of parameters and operations, making the model more lightweight and suitable for deployment on resource-constrained devices.

The real-time capabilities of QuartzNet make it a valuable choice for on-device ASR tasks, ensuring a seamless user experience across a wide range of applications. Whether it's enabling voice-controlled devices, virtual assistants, transcription services, or voice-enabled applications in various industries, QuartzNet showcases its prowess in real-time speech recognition.

QuartzNet derives its foundation from the Jasper model, integrating separable convolutions and larger filters. Remarkably, QuartzNet achieves performance similar to Jasper while boasting an order of magnitude fewer parameters. As part of the QuartzNet family of models, they are denoted as QuartzNet_[BxR], with B representing the number of blocks, and R indicating the number of convolutional sub-blocks within each block [7]. Each sub-block includes a 1-D separable convolution, batch normalization, ReLU activation, and dropout, showcasing the model's architectural finesse and effectiveness [7].

Furthermore, Nvidia NeMo's Citrinet takes ASR capabilities to even greater heights. As an evolution of QuartzNet, Citrinet extends its capabilities by incorporating elements from ContextNet, Word Piece tokenization, and a non-autoregressive CTC-based decoding scheme [8]. Citrinet combines the strengths of both convolutional neural networks (CNNs) and transformer-based architectures, bringing together powerful audio processing and long-range dependency capture [8].

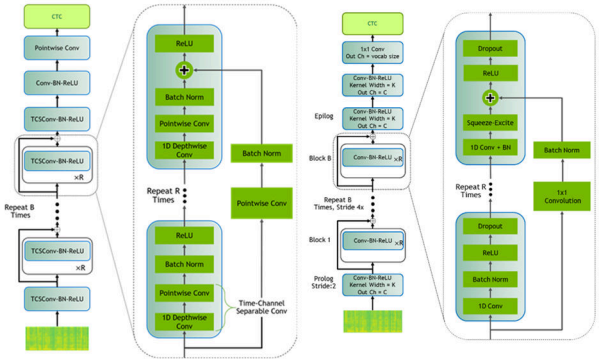

Figure 3. Architectures of QuartzNet (left) [7] and Citrinet (right) [8]

By leveraging CNNs, Citrinet efficiently processes audio and extracts essential features, while the transformer architecture empowers the model to capture long-range dependencies within speech signals, especially when used in tandem with a language model [8]. The inclusion of the transformer component bestows Citrinet with the ability to consider a broader context, making it exceptionally proficient in tasks where context is critical for accurate recognition. This combination yields superior performance in ASR tasks, particularly in scenarios where comprehending context is essential for precise interpretation of speech.

Like NeMo QuartzNet, Citrinet is engineered for efficiency, rendering it ideal for on-device and real-time ASR

applications. Its training on extensive datasets further equips Citrinet with the capacity to recognize speech across multiple languages and adapt to varying accents and speaking styles, making it a versatile and reliable ASR solution for various practical applications.

IV. Training and results

The dataset used in this study consists of a total duration of 34.84 hours, divided into three sets: training (20.92 hours), testing (6.96 hours), and development (6.96 hours), following a 60-20-20% ratio.

The experiments for training and evaluating the acoustic models were conducted on a machine with a 64-bit Linux OS (Ubuntu 20.04 LTS) and an Intel Core i7-9700 CPU clocked at 3 GHz. The machine is equipped with two NVIDIA GPUs: NVIDIA GeForce GTX 1660 TUF and NVIDIA GeForce GTX 1650, with a total of 10GB of memory. The models were trained with identical hyperparameters, including a learning rate of 0.001, 100 epochs, and a batch size of 4.

For transfer learning, pre-trained English models were used, and the transfer learning feature was applied to these models.

Data augmentation was employed to enhance the dataset size and introduce speech variations for the models. Various augmentation techniques were applied, including sample domain augmentation, spectrogram domain augmentation, and multi-domain augmentation for DeepSpeech. For QuartzNet and Citrinet, "SpecAugment" [13] and "SpecCutout" [14] techniques were used to augment the spectrograms, effectively diversifying the training data.

In a unique approach, the language model was trained using a distinctive corpus of text collected from Armenian news websites, covering diverse subjects like lifestyle, music, culture, politics, and sports. The Python console application effectively scraped and normalized the text, resulting in a comprehensive dataset of 9,539,350 sentences, providing accuracy and versatility in ASR tasks.

The acoustic model's efficiency and accuracy were measured using two fundamental metrics: CER (character error rate) and WER (word error rate). CER assesses the accuracy of the acoustic model in recognizing individual characters, while WER evaluates the accuracy of the language model in recognizing words.

After 100 epochs of training, the models were evaluated both with and without language models.

The Deepspeech model has a size of 180 megabytes with 2048 hidden layers. The QuartzNet model has 18.9 million trainable parameters and a size of 37.87 megabytes. The Citrinet model includes 36.4 million trainable parameters and a built-in tokenizer, with an estimated size of 72.73 MB. Additionally, the language model used for evaluation has a size of 170 megabytes.

For a comprehensive overview of the results, including training and testing with and without language models, please refer to Table 1.

Model	Size (MB)	WER/CER	WER/CER with LM
DeepSpeech	180	0.6302/0.1435	0.3231/0.1259
QuartzNet	37.87	0.3351/0.1537	0.2676/0.1092
Citrinet	72.73	0.1941/0.8266	0.0869/0.0372

Table 1. Results

V. Conclusion

This research involved a thorough comparison of three state-of-the-art ASR models for Armenian speech recognition: DeepSpeech, Nvidia NeMo QuartzNet, and Citrinet. Following a 100-epoch training phase, the models underwent evaluation both with and without the incorporation of a language model.

DeepSpeech achieved a standalone WER of 63.02% and a CER of 14.35% with a model size of 180 MB and 2048 hidden layers. The incorporation of a language model significantly improved performance, reducing WER to 32.31% and CER to 12.59%.

QuartzNet exhibited impressive performance on its own, achieving a WER of 33.51% and a CER of 15.37%. With the integration of a language model, the WER improved slightly to 26.76%, and the CER reached 10.92%. Notably, QuartzNet's compact model size of only 37.87 MB further adds to its appeal.

Citrinet exhibited excellent standalone performance with a WER of 19.41% and a CER of 8.26%. With a model size of 72.73 MB (including a built-in tokenizer), integrating a language model further improved the results to a remarkable WER of 8.69% and a CER of 3.72%.

In summary, Citrinet performed exceptionally well, achieving the lowest error rates, and DeepSpeech showed significant improvement with a language model. QuartzNet's competitive performance was limited by its inability to integrate a language model.

These findings provide valuable insights for Armenian speech recognition, helping researchers and practitioners choose suitable ASR models based on their specific requirements and language context. Additionally, the model size comparison aids in decision-making for resource-constrained environments. This research contributes to advancing ASR technology for the Armenian language and diverse speech corpora.

References

[1] Acoustic model, In Wikipedia, [online]. https://en.wikipedia.org/wiki/Acoustic_model.

[2] Language model, In Wikipedia, [online]. https://en.wikipedia.org/wiki/Language_model.

[3] K. Heafield, “KenLM: Faster and Smaller Language Model Queries”, *WMT at EMNLP*, pp. 30-31, July 2011.

[4] M. Siu and M. Ostendorf, “Variable n-grams and extensions for conversational speech language modeling”, *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 63–75, 2000.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Y. Ng, “Deep Speech: Scaling up end-to-end speech recognition”, *arXiv preprint arXiv:1412.5567*, 19 December 2014.

[6] Mozilla, “Project DeepSpeech”, 2021, [online]. https://github.com/mozilla/ DeepSpeech.

[7] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions”, *arXiv preprint arXiv:1910.10261*, 2019.

[8] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi, B. Ginsburg, “Citrinet: Closing the Gap between Non-Autoregressive and Autoregressive End-to-End Models for Automatic Speech Recognition”, *arXiv preprint arXiv:2104.01721*, 2021.

[9] V. H. Baghdasaryan, “ArmSpeech: Armenian Spoken Language Corpus”, *International Journal of Scientific Advances (IJSCIA)*, vol. 3, no. 3, pp. 454-459, May-Jun 2022.

[10] V. H. Baghdasaryan, “Extended ArmSpeech: Armenian Spoken Language Corpus”, *International Journal of Scientific Advances (IJSCIA)*, vol. 3, no. 4, pp. 573-576, Jul-Aug 2022.

[11] S. Chakmakjian and H. Dolatian, “Speech corpus of Armenian question-answer dialogues”, 2022, DOI: 10.5281/zenodo.7088365.

[12] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, A. Bapna, “FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech”, *IEEE Spoken Language Technology Workshop (SLT)*, pp. 798-805, 2022.

[13] D. S. Park, W. Chan, Y. Zhang, CC. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”, *arXiv preprint arXiv:1904.08779*, 2019.

[14] T. DeVries, G. W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout”, *arXiv preprint arXiv:1708.04552*, 2017.