Bi-dialectal ASR of Armenian from Naturalistic and Read Speech

Arthur Malajyan[1], Victoria Khurshudyan[2], Karen Avetisyan[3],
Hossep Dolatian[4], Damien Nouvel[5]
[1,3]Russian-Armenian University, [2]INALCO/SEDYL/CNRS , [4]Stony Brook University, INALCO
malajyanarthur@ispras.ru, victoria.khurshudyan@inalco.fr, karavet@ispras.ru,
hossep.dolatian@alumni.stonybrook.edu, damien.nouvel@inalco.fr

Abstract

The paper explores the development of Automatic Speech Recognition (ASR) models for Armenian, by using data from two standard dialects (Eastern Armenian and Western Armenian). The goal is to develop a joint bi-variational model. We achieve **state-of-the-art** results. Results from our ASR experiments demonstrate the impact of dataset selection and data volume on model performance. The study reveals limited transferability between dialects, although integrating datasets from both dialects enhances overall performance. The paper underscores the importance of dataset diversity and volume in ASR model training for under-resourced languages like Armenian.

Keywords: Armenian, ASR, oral corpus, speech corpus, dialect, naturalistic speech corpus

1. Introduction

Armenian is an Indo-European language with two standard dialects – Standard Eastern Armenian and Standard Western Armenian – along with dozens of non-standard dialects. Eastern Armenian is the official language of Armenia, and is spoken by the Eastern Armenian diaspora in Russia, Georgia, Iran, and elsewhere. Western Armenian developed in the Ottoman Empire, and it became a diasporic dialect following the Armenian Genocide.

Armenian is generally considered a low-resource language (Megerdoomian, 2009; Vidal-Gorène et al., 2020). Though Eastern Armenian has more resources than Western Armenian (discussed in Dolatian et al., 2022). In terms of speech resources, Eastern Armenian has the Eastern Armenian National Corpus (Khurshudian et al., 2009; Khurshudyan et al., 2022), which includes an oral corpus. There are some working ASR models for Eastern Armenian: Armspeech,[1] ican24,[2] arampacha.[3] These models have generally not been tested for their performance with respect to Western Armenian. See discussion on bi-dialect Armenian ASR in Chakmakjian and Wang (2022).

The present study is conducted as part of the project *DALiH*, or *Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus and Data Processing* in collaboration with the *Center of Advanced Software Technologies* at the Russian-Armenian University.[4] The DALiH project seeks to set up a comprehensive linguistic digital platform for both diachronic and synchronic varieties of the Armenian language. This platform aims to provide open-access and open-source resources, including grammatically annotated corpora, along with various annotation tools such as dictionaries, datasets, and annotation models based on different approaches.

The project also aims to incorporate oral corpora, representing standard Western and Eastern Armenian, as well as several modern dialects. One of the key objectives of the project is to develop Automatic Speech Recognition (ASR) models for Eastern and Western based on text-speech aligned oral corpora. The automatic alignment task itself presents a significant challenge that needs to be addressed. Current advancements in NLP offer promising opportunities not only to utilize NLP resources from well-resourced languages for under-resourced ones but also to re-purpose existing resources for various linguistic varieties within a target language, rather than creating new resources from scratch. Consequently, this research aims to explore the development of a joint bi-variational model for Eastern and Western Armenian, potentially offering more efficient solutions for under-resourced languages in a multivariational context.

This paper is organized as follows. We provide background information (§2) on Armenian phonology, phonetics, and orthography, and on Armenian ASR. We describe our ASR experiments in §3. We conclude and discuss the results in §4.

[1] https://pypi.org/project/armspeech/
[2] https://hayq.ican24.net/asr/index.php
[3] https://huggingface.co/arampacha/whisper-large-hy-2
[4] The DALiH project is funded by French National Research Agency ANR-21-CE38-0006.

SIGUL2024 Workshop, pages 227–236

Table 1: Comparison of laryngeal contrasts for stops and affricates

	Eastern	Western			
		Turkey	Lebanon	USA	
⟨բ⟩ ⟨բան⟩	**b**ɑn	**p**^{h}ɑn	**p**ɑn	**p**^{h}ɑn	'thing'
⟨պ⟩ ⟨պահ⟩	**p**ɑh	**b**ɑh	**b**ɑh	**p**ɑh	'period'
⟨փ⟩ ⟨փայլ⟩	**p**^{h}ɑjl	**p**^{h}ɑjl	**p**ɑjl	**p**^{h}ɑjl	'shine'

2. Background

2.1. Linguistic Differences in Armenian

When designing multi-variational or multi-dialectal ASR models, one should keep in mind major phonological and orthographic differences between dialects.

A non-trivial phonological difference between the two standard Armenian varieties (and many other non-standard dialects) is differences in the laryngeal quality of stops and affricates (Table 1).

In Eastern Armenian, stops and affricates have a phonemic contrast in being voiced, voiceless unaspirated, or voiceless aspirated. The three phonemic categories are represented by distinct orthographic letters (Vaux, 1998; Hacopian, 2003; Seyfarth and Garellek, 2018). Yet other dialects like Western Armenian (and its regional subdialects) have simplified or altered the voicing system, while still keeping the orthographic system. Western Armenian specifically has simplified the three-way laryngeal contrast into a two-way one. For example the letter ⟨պ⟩ marks a phonemically voiceless unaspirated stop /p/ in Eastern Armenian, but in Western Armenian, it is a voiced stop /b/ though the pronunciation can vary by region from [b] to [p] (Kelly and Keshishian, 2021; Seyfarth et al., 2023).

Armenian orthography has two distinct letters to represent rhotics: ⟨ռ, ր⟩. The letter ⟨ր⟩ marks an alveolar flap /ɾ/ in Eastern Armenian and Western Armenian. The letter ⟨ռ⟩ is a trill /r/ in Eastern but a flap /ɾ/ in Western. the voiced alveolar trill /r/ ⟨ռ⟩ and the alveolar tap /ɾ/ . Some dialects that we plan to incorporate in the future add further rhotic distinctions. For example the flap /ɾ/ ⟨ր⟩ is replaced with an approximant /ɻ/ in Iranian Armenian (Dolatian et al., 2023).

Various other phonetic discrepancies between the dialects arise from different factors, including areal contact-induced phonetic changes. Notable examples in Eastern Armenian include the optional realization of voiceless unaspirated stops like /k/ as ejectives [k'] (e.g., կապիկ [kɑpik, k'ɑpik'] 'monkey'), the tendency to palatalize certain consonants because of Russian influence (e.g., սուբյեկտիվ [subjektiv, subjekt^{j}iv] 'subjective'), and the possible rounding of low back vowel /ɑ/ as [ɒ] because of Persian influence, often in Iranian Eastern Armenian (Dolatian et al., 2023). The Eastern glide-vowel sequence /ju/ has multiple possible pronunciations in Western Armenian ([ʏ, uj], such as how the word 'flour' is Eastern Armenian [ɑljuɾ] ⟨ալյուր⟩ but Western [ɑlʏɾ, ɑlujɾ] ⟨ալիւր⟩.

An orthographic difference is that until the 1920s, both Western and Eastern Armenian were written with the same spelling system in the Armenian script. But during the Soviet Union, various spelling reforms were made for the Eastern Armenian community in modern-day Armenia and Russia, but not for Eastern Armenian communities in Iran nor for Western Armenian communities (Sanjian, 1996). For example, the word 'love' is pronounced [seɾ] in both dialects. The traditional spelling (as used by Western Armenian and *Iranian* Eastern Armenian) is ⟨սէր⟩ with the letter ⟨է⟩ for /e/; while the reformed spelling for Eastern Armenian is ⟨սեր⟩ with the letter ⟨ե⟩ for /e/.

2.2. Background on ASR

Both Armenian dialects have a rich written tradition with ample texts. But in contrast to written materials, oral data in Armenian is seldom accessible for research purposes. This is the case for Eastern Armenian, Western Armenian, and non-standard dialects. This scarcity of source data indirectly contributes to the shortage of ASR models. In recent years, several projects have endeavored to develop ASR models for Eastern Armenian (Google Translate,[5] the Public initiative for national acceleration or Ազգային արագացման հանրային նախաձեռնություն (ican24),[6] Mozilla Common Voice,[7] Sonix,[8], HindiTyping,[9] wav2vec 2.0[10]).

The main challenge of ASR model designing is the training and evaluation of one or several ASR models for the Armenian varieties. Most state-of-the-art ASR tools require hundreds or thousands of transcribed data as the training dataset, but the recent rise of interest for low- and medium-resource languages such as Armenian pushed some of them to address the challenge to offer models that require a restricted or limited transcribed dataset (i.e., few-shot learning).

[5] https://translate.google.com/?hl=hy&sl=hy&tl=la&op=translate

[6] https://arm.ican24.net/demoasrv4.html

[7] https://pontoon.mozilla.org/hy-AM/common-voice/

[8] https://sonix.ai/languages/transcribe-armenian-audio

[9] https://hindityping.info/speech-to-text/armenian/

[10] https://huggingface.co/infinitejoy/wav2vec2-large-xls-r-300m-armenian

Among those tools, Whisper (Radford et al., 2022) and SeamlessM4T (Communication et al., 2023) models are large multilingual models trained on datasets consisting of more than 100 languages. Both Whisper and SeamlessM4T have been trained on a diverse dataset, making it robust and versatile for transcription tasks. They are particularly noted for their high accuracy and the ability to recognize context, which helps in providing more accurate transcriptions. Both of them are also achieving state-of-the-art result for many low- and under-resourced languages. By using these models, new data can be added at each iteration and help speed up manual correction.

Once the training set reaches a substantial size, other approaches will be possible to be tested, including transfer learning from a high-resource language, as studies showed that they give good results if fine-tuned with at least 20 hrs (Mohamud et al., 2021) or 35 hours (Hjortnaes et al., 2020) of transcribed data of the target language. Interestingly, Mohamud et al. (2021) showed that applying a self-supervising model trained on a given language as the backbone produces "indistinguishable results on languages originating from the same family."

3. ASR Methodology and Results

3.1. Data

Our speech data was taken from different sources summarized in Table 2. We had more data from Eastern Armenian than Western. Some data was read speech, and some was naturalistic speech. Each data source was given a code.

Table 2 summarizes the amount of hours used across the training, development, and test sets.

3.1.1. Common Voice (CV)

Common Voice (Ardila et al., 2019)[11] is a volunteer-driven initiative launched by Mozilla. It aims at building an open-source database for speech recognition applications for more than 100 languages. This project relies on contributions from volunteers who record examples of speech and evaluate the recordings submitted by others. Specifically for the Armenian language the volunteers are given sentences from the Eastern Armenian Wikipedia and their task is to pronounce them. Most of the recordings were in Eastern Armenian. We used the 16.1 version of Common Voice.

[11] https://pontoon.mozilla.org/hy-AM/common-voice/

3.1.2. Google Fleurs (GF)

Google Fleurs (Conneau et al., 2022)[12] is a comprehensive dataset for speech recognition research that encompasses parallel speech data in 102 languages. Fleurs is an open-source dataset that includes nearly 12 hours per language for over 100 languages. It is based on Wikipedia sentences. Each sentence for each language was pronounced by 3 different native speakers. The Armenian data is in Eastern Armenian.

3.1.3. Eastern Armenian National Corpus (EA)

The EANC[13] contains approximately 110 million tokens of Eastern Armenian data spanning from the mid-19th century to the present (Khurshudian et al., 2009; Khurshudyan et al., 2022). It includes written and oral data, with the texts and transcripts annotated grammatically (POS-tagging, full-fledged morphological and semantic tagging) and metatextually. The oral sub-corpus consists of spontaneous dialogues, polylogs, task-oriented narratives, TV talk shows, movies, and other recordings across various subgenres. The oral data (nearly 3 million tokens, 350 hrs) were compiled and transcribed as part of the EANC initiative (Table 3).

The EANC oral subcorpus data that we used is approximately 6 hours of authentic oral data, primarily consisting of interviews and talk shows. The data was constrained in order to ensure comparability between WA and EA datasets, given that the available data for Western Armenian amounted to approximately 6 hours. This data was collected from various television media outlets in Armenia between 2006 and 2009. The data underwent pre-alignment, conversion to Praat TextGrid format, and manual correction. The alignment process was primarily semi-automated, involving the initial use of a forced alignment tool to preprocess the data, followed by manual realignment by experts from the DALiH project. Forced alignment consists in matching a given transcript to the sound, commonly on the word level, and sometimes with the help of automatic phoneme identification. Within the DALiH project, the tool aeneas[14] was employed, as it utilizes a text-to-speech engine specifically developed (naively) for Armenian (both Eastern and Western), with the option for fine-tuning.

[12] https://huggingface.co/datasets/google/fleurs

[13] http://www.eanc.net/

[14] https://github.com/readbeyond/aeneas

Table 2: Speech data used and the size of the data

Code	Source	Dialect	Speech type	Train	Dev	Test
CV	Common Voice	Eastern	Read	5,5 hr.	4 hr.	4,5 hr.
GF	Google Fleurs	Eastern	Read	10,5 hr.	1,2 hr.	3 hr.
EA	EANC	Eastern	Naturalistic	5,8 hr.	0,5 hr.	0,5 hr.
WA	ReRooted	Western	Naturalistic	5,8 hr.	0,5 hr.	0,5 hr.

Table 3: EANC Oral Data Composition (as of February 2024)

Oral discourse	# tokens	% EANC	# of docs
Spontaneous discourse	1 029 646	29,6%	208
Public discourse	1 933 899	55,6%	543
Task-oriented discourse	70 010	2,0%	22
Online communication	442 399	12,7%	1
Total	3 475 954	100%	774

3.1.4. ReRooted (WA)

The above sources are for Eastern Armenian. For Western Armenian, we used the ReRooted corpus.[15] ReRooted is an oral history of refugee testimonials by over 100 Syrian Armenians who fled the Syrian Civil War (Baghdassarian and Broidy, 2018). As of Jan 31 2024, the corpus has 75hrs of WA speech, along with time-aligned captions. A 6hr subset of those testimonies have been converted to Praat TextGrids and manually corrected (about 6hr with 9 speakers). We use those 6hrs (Dolatian, 2024).

3.2. Models

We were inspired by the novel multilingual big speech recognition models that achieve SOTA results from out-of-the-box systems for different low-resource languages. So we decided to use the different Whisper models released by OpenAI and the different Seamless models released by Meta. These models are multilingual. They have been trained on Armenian language data as well. The subsequent sections describe the utilized models and provide a detailed description of the architectures of the aforementioned models.

3.2.1. Whisper Large v1

Whisper Large v1[16] is a Transformer-based encoder-decoder, sequence-to-sequence model. This architecture not only transcribes speech but also employs the decoder as a language model to enhance language comprehension and minimize grammatical errors. Whisper v1 was trained on 680k hours of annotated speech data annotated with large-scale weak supervision. This version of Whisper demonstrates adaptability in processing both monolingual and multilingual datasets. While monolingual training primarily focuses on speech recognition tasks, the multilingual aspect also has speech translation capabilities.

3.2.2. Whisper Large v2

Whisper Large v2[17] shares the same architecture as Whisper v1. However, the key difference lies in the training regimen, where the number of training epochs for Whisper v2 was increased by 2.5 times, incorporating techniques such as SpecAugment, stochastic depth, and BPE dropout for regularization purposes.

3.2.3. Whisper Large v3

Whisper Large v3[18] retains the architecture of its predecessors while introducing certain enhancements. Notably, the input representation now utilizes 128 Mel frequency bins instead of the previous 80, and a new language token for Cantonese has been incorporated. Whisper v3 was trained on a combined dataset comprising 1 million hours of weakly labeled audio and 4 million hours of pseudolabeled audio, collected using Whisper large-v2. The training process spanned 2.0 epochs over this amalgamated dataset, resulting in further improvements in performance and versatility.

3.2.4. SeamlessM4T v1

SeamlessM4T[19] (Massively Multilingual & Multimodal Machine Translation) is a multitask model based on the multitask UnitY (Inaguma et al., 2023) model architecture. It is designed to directly generate translated text and speech, encompassing various translation tasks including automatic speech recognition, text-to-text, text-to-speech, speech-to-text, and speech-to-speech translations.

To construct this model, 1 million hours of speech audio data were utilized to train self-

[15] https://www.rerooted.org/

[16] https://huggingface.co/openai/whisper-large

[17] https://huggingface.co/openai/whisper-large-v2

[18] https://huggingface.co/openai/whisper-large-v3

[19] https://huggingface.co/facebook/seamless-m4t-large

supervised speech representation. Additionally, a corpus of aligned speech translations (470,000 hours) was employed. In contrast to Whisper, this approach facilitated the development of the first multilingual system capable of bidirectional translation involving English for both speech and text.

3.2.5. SeamlessM4T v2

SeamlessM4T v2[20] is built upon the UnitY2 model architecture, setting it apart from its predecessor, SeamlessM4T v1. Unlike v1, the text-to-unit decoder component in v2 is non-autoregressive, allowing for adaptation to streaming scenarios. Furthermore, v2 incorporates an additional 114,800 hours of speech and text alignments, supplementing the existing dataset. This augmentation not only expands the total hours but also broadens language coverage from 37 to 76 languages. Moreover, v2 can preserve vocal styles and prosody during translation.

3.2.6. Dedicated Armenian Models

ArmSpeech is an Armenian speech-to-text library utilizing Coqui STT.[21] The model is a recurrent neural network (RNN) with five layers of hidden units, and it has been trained using the ArmSpeech dataset (Baghdasaryan, 2022) consisting of 15,7 hours. The acoustic model collaborates with the language model to enhance the accuracy of predictions. The language model is based on the KenLM Language Model Toolkit library.[22] **Arampacha** is a model available on Huggingface[23] and is based on the Whisper-large-v2 model after being fine-tuned with Common Voice v11.0[24]. The only information known about the **ican24** is that it is a model based on Vosk v17.0.[25]

3.3. Experiments

Two types of experiments have been conducted based on fine-tuning the different models using different types of data (Eastern only vs. Western only vs. bi-dialectal, naturalistic speech vs. read speech vs. both).

In the first experiment, we aimed to mimic the scenario where there already exists a pretrained model for the Armenian language, and we sought to fine-tune it using specific datasets. Initially, we fine-tuned the Whisper and Seamless models using the Common Voice (CV) and Google Fleurs (GF) datasets. These models were thus fine-tuned using read speech. Subsequently, this fine-tuned model underwent another round of tuning on the naturalistic speech datasets: Eastern Armenian from EANC (EA) and Western Armenian from ReRooted (WA) datasets (naturalistic speech). These experiments were conducted to assess whether a model trained on the EA dataset could effectively perform speech recognition for WA and vice versa. Furthermore, we also fine-tuned the models using combined EA + WA datasets to aim for the highest overall performance across all tests.

For the second type of experiments, we started tuning the models from the checkpoints of the Whisper and Seamless models. Initially, we tuned them using only data from either the EA or WA datasets (naturalistic speech). These experiments were carried out to investigate the transferability of knowledge between these two dialects. Additionally, we separately fine-tuned the models using combined EA + WA and CV + GF + EA + WA datasets to maximize results and observe the impact of increasing the volume of data.

The final set of experiment scenarios is 9. They are outlined as follows (-> denotes fine-tuning):

1. Out-of-the-Box -> CV + GF
2. Out-of-the-Box -> CV + GF -> EA
3. Out-of-the-Box -> CV + GF -> WA
4. Out-of-the-Box -> CV + GF -> EA + WA
5. Out-of-the-Box -> EA
6. Out-of-the-Box -> WA
7. Out-of-the-Box -> EA+WA
8. Out-of-the-Box -> CV + GF + EA + WA
9. Out-of-the-Box

The models were trained for numerous epochs until they reached a plateau in terms of metrics. We used different hyperparams for Whisper and Seamless models. For the Seamless models the batch size = 4, learning rate = 1e-6, max epoch number = 20. For the Whisper models the batch size = 4, learning rate = 1e-5, max epoch number = 20, we also freezed Whisper's encoder part. These metrics were computed on four development sets, with each set corresponding to a different type of training data. Subsequently, the average of these four results was calculated. For the final results, we selected the model from the epoch with the best results on the average of the development sets.

as

3.4. Results

After fine-tuning, the models were tested tested on all four datasets corresponding to the training and

[20] https://huggingface.co/facebook/seamless-m4t-v2-large

[21] https://stt.readthedocs.io/en/latest/

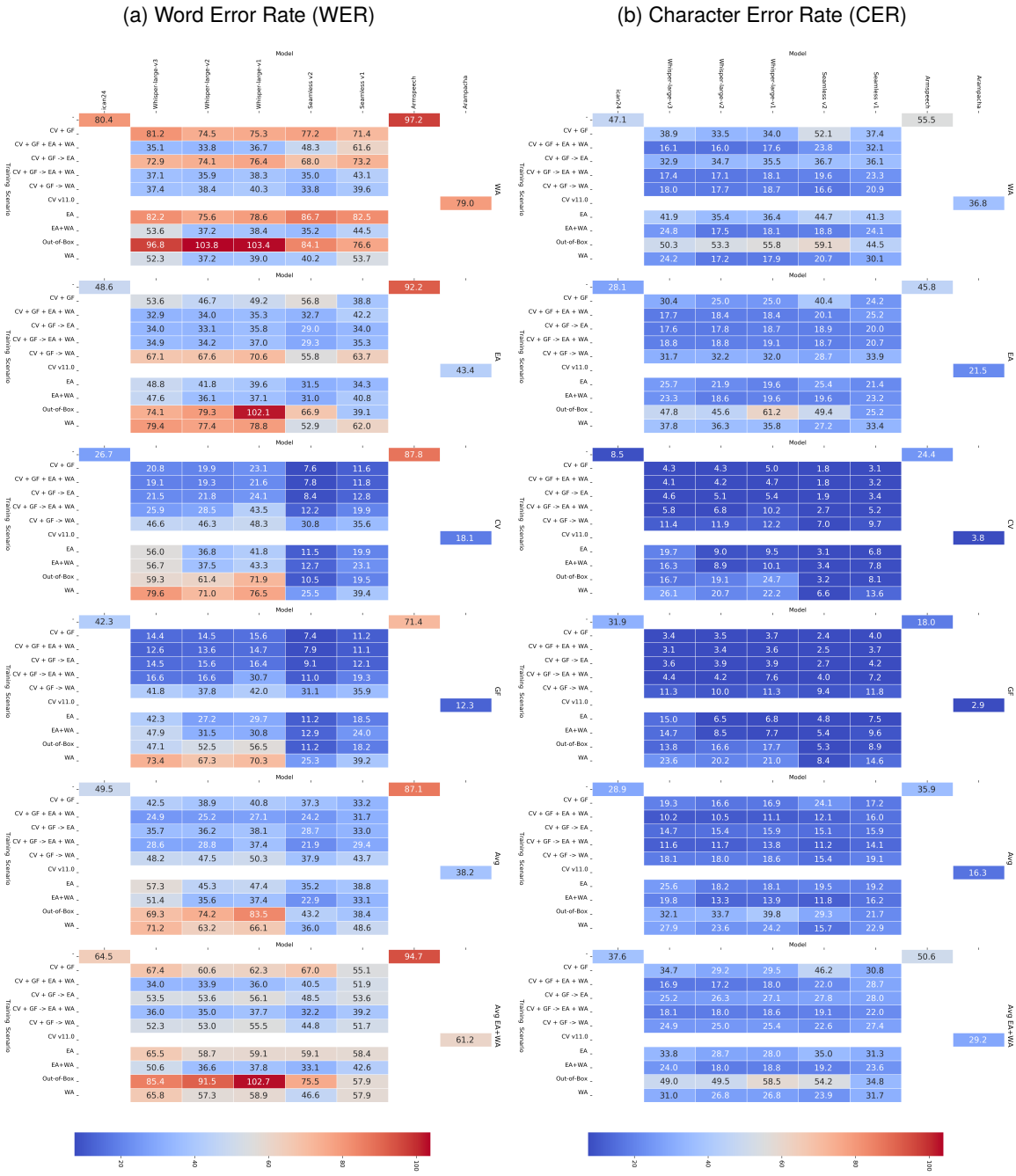
[22] https://kheafield.com/code/kenlm/

[23] https://huggingface.co/

[24] https://commonvoice.mozilla.org/en/datasets

[25] https://alphacephei.com/vosk/

Figure 1: Results (WER and CER) from testing the models on test sets, after fine-tuning with different scenarios.



development sets. Figure 1 reports the Word Error Rate (WER) and Character Error Rate (CER) on the test sets.[26] We likewise tested dedicated Armenian models (Armspeech, ican24, Arampacha).

The results clearly demonstrate that incorporating a specific dataset within the training set leads to an improvement in metrics for the corresponding test sets. This means that if a model was trained on the CV training data, then it did well on the CV test data.

Moreover, augmenting the volume of data used for model training generally enhances results on average.

For Whisper-based models, there is a notable contrast between the WA-trained model and the one trained solely on EA data. Specifically, the EA-trained model shows increased metrics for both the CV and GF test sets compared to the Out-of-

[26]We thank Chahan Vidal-Gorène for help in making these figures.

Table 4: Models that achieved best WER and CER results on different test sets

	Model	Training Scenario	WER	CER
Best WA model	Whisper-large-v2	CV + GF + EA + WA	33,8	16,0
Best EA model	Seamless v2	CV + GF ->EA	29,0	18,9
Best CV model	Seamless v2	CV + GF	7,6	1,8
Best GF model			7,4	2,4
Best EA and WA Avg. model	Seamless v2	CV + GF ->EA + WA	32,2	19,1
Best all tests Avg. model			21,9	11,2

the-Box scenario. This phenomenon could be attributed to the fact that the CV and GF datasets predominantly consist of Eastern Armenian speech. Conversely, for Seamless models, the results are largely comparable to the Out-of-the-Box scenario.

Overall, the results indicate that using open-source datasets alone does not adequately address the challenge of deploying models trained on datasets from other domains. For instance, models fine-tuned on CV and GF datasets (which are read speech) exhibit poor performance on EA and WA tests (which are naturalistic speech).

The language (dialect) transferability is notably limited. Models trained on EA performed poorly on WA tests, and vice versa. However, despite this limitation, the results showed improvement compared to the Out-of-the-Box scenarios. This suggests that datasets from different dialects do provide some assistance in the task of speech recognition for other dialects/varieties. Nevertheless, achieving high results for specific dialects necessitates access to datasets specifically tailored to those dialects.

Another notable observation is that EA and WA datasets can mutually benefit each other. Whisper models trained on a combined EA + WA dataset demonstrated superior performance on both EA and WA test sets compared to models trained solely on EA or WA data.

The achieved results surpass those of the Out-of-the-Box models for both Eastern and Western Armenian. However, the decision on whether it is more advantageous to utilize a pre-trained model and fine-tune it or train from scratch with the entire dataset starting from a multilingual pre-trained checkpoint varies from model to model.

In Table 4, we present the best results obtained for each of the test sets, as well as the best average results for EA and WA individually, along with the average results for all four test sets. Notably, we achieved a WER of nearly 30% for both EA and WA test sets, and exceptionally high results for the GF and CV sets, reaching approximately 7.5% WER.

Table 5 showcases the best results achieved by each model, juxtaposed with the existing results for Armenian language models. Notably, Seamless v2 attained the best WER results, while Whisper v3 excelled in terms of CER.

Table 5: The best test-averaged results achieved by each model

Model	Training Scenario	WER	CER
Whisper-large-v1	CV + GF + EA + WA	27,1	11,1
Whisper-large-v2	CV + GF + EA + WA	25,2	10,5
Whisper-large-v3	CV + GF + EA + WA	24,9	**10,2**
Seamless v1	CV + GF ->EA + WA	29,4	14,1
Seamless v2	CV + GF ->EA + WA	**21,9**	11,2
ArmSpeech	ArmSpeech	87,1	35,9
ican24	-	49,5	28,9
Arampacha	CV v11.0	38,2	16,3

3.5. Error Analysis

We performed a comparative analysis of the best two models (Table 5) to identify the types of errors that each model made and to determine their respective strengths under various conditions. To facilitate this comparison, transcriptions from both models across all tests were examined. Instances where one model performed well and the other did not were particularly examined.

The Seamless v2 model (CV + GF ->EA + WA) sometimes misinterpreted Eastern Armenian speech as Western Armenian. This misinterpretation involved using different spelling systems (Table 6a; such as using Classical orthography instead of Reformed orthography) or not transcribing an entire suffix (Table 6b).

In contrast, Whisper v3 (CV + GF -> EA + WA) demonstrated difficulties in transcribing Western Armenian speech. In (c), the sentence 'we got' uses a periphrastic construction /ɑɾeɾ ejiŋkʰ/ ‹առեր էինք› that only exists in Western Armenian, not Eastern. Yet it transcribed it as a non-existing word /ɑɾɑjiŋkʰ/

The model sometimes resorted to abbreviations (d) or omitted parts of the audio (e). For (d), it abbreviated the word 'with kilograms', while (e) omitted entire words.

In sum, Seamless v2 demonstrates a higher accuracy in transcribing Western Armenian texts compared to Whisper v3. However, it occasionally translates dialects, converting Eastern Armenian into Western Armenian. Although Whisper v3 exhibits fewer of these specific errors, it tends to

Table 6: Types of errors made by the best-performing models

	Model	Audio (IPA)	Correct transcription	Model's incorrect transcription	Pronunciation of incorrect transcription
(a)	Seamless v2	/ɑmɑrvɑ/	ամառվա	ամառուայ	/ɑmɑrvɑ/
(b)	Seamless v2	/tʰɑlɑnvet͡sʰ/	թալանվեց	թալանուեցաւ	/tʰɑlɑnvet͡sʰɑv/
(c)	Whisper v3	/ɑrer ejiŋkʰ/	առեր էինք	առայինք	/ɑrɑjiŋkʰ/
(d)	Whisper v3	/kilogɾɑmov/	կիլոգրամով	կգով	/kilogɾɑmov/
(e)	Whisper v3	/t͡sʰənt͡sʰumə meʁɾɑmisi pʰuln ɑveli/	ցնցումը մեղրամիսի փուլն ավելի	ցնցումը ավելի	t͡sʰənt͡sʰumə ɑveli

leave out parts of the audio or resort to abbreviations in the transcription.

4. Conclusion and Future Perspectives

Our experiments have provided valuable insights into the effectiveness of various training strategies and datasets for speech recognition models in Eastern and Western Armenian dialects. Key findings include:

- The incorporation of specific datasets into the training process leads to improvements in test set metrics, underscoring the importance of dataset selection in model training.
- Increasing the amount of data generally enhances model performance, highlighting the crucial role of data quantity in training models effectively.
- Whisper-based models trained exclusively on Eastern Armenian data demonstrated improved performance on test sets such as Common Voice and Google Fleurs, likely due to the prevalence of Eastern Armenian speech in these datasets.
- The language/variety transferability is limited, with models trained on Eastern Armenian showing poor performance on Western Armenian tests and vice versa. However, integrating datasets from different varieties can still mutually enhance model performance for both dialects.
- Our results surpass Out-of-the-Box models, with WER reaching nearly 30% for both Eastern and Western Armenian test sets and approximately 7.5% for Common Voice and Google Fleurs sets.
- Surprisingly, multi-lingual models like Whisper and Seamless outperformed the monolingual models that were solely trained on Armenian like ArmSpeech and ican24.

The analysis of the results clearly shows the development of **state-of-the-art** models for both Western and Eastern Armenian languages. Moreover, beyond the Armenian dialectal variations, our findings serve as a valuable case study for the development of ASR models, particularly in the context of low-resource languages in a multivariational context.

A potential avenue for future research would involve increasing the **amount of data** in both Eastern and Western varieties, as well as other dialects, taking into account data accessibility, to assess the impact on model training efficiency based on target language and variety-based data.

Another aspect to explore would be the quality of the data, with the hypothesis that more **naturalistic** data may require less volume. Many existing models rely on somewhat artificial data sources, such as readings of written texts like audiobooks or Wikipedia articles. It is thus interesting to increase the amount of naturalistic data instead of read speech.

Given that the DALiH project encompasses a comprehensive approach to processing Armenian language variation across various NLP aspects, it would be intriguing to compare the efficiency of transferability in annotation and automatic speech recognition processing. The hypothesis here is that annotation transferability may be higher than ASR transferability, as the written-orthographic layer can potentially bridge more of the differences between varieties than phonemic or phonetic differences.

Another perspective within the DALiH project could entail assessing how a phonetic dictionary impacts ASR performance. This endeavor is in line with the project's overarching goal of integrating linguistic principles with NLP methodologies, aiming to elevate the role of linguistics within the NLP domain, particularly in a research context, despite the perceived idealism associated with such an endeavor.

5. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer,

Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Varuzhan H Baghdasaryan. 2022. Armspeech: Armenian spoken language corpus. *International Journal of Scientific Advances (IJSCIA)*, 3(3):454–459.

Anoush Baghdassarian and Lauren Broidy. 2018. Documenting 100 years of displacement among Syrian-Armenians: An interview with Anoush Baghdassarian conducted by Lauren Broidy. *Review of Middle East Studies*, 52(2):334–343.

Samuel Chakmakjian and Ilaine Wang. 2022. Towards a unified ASR system for the Armenian standards. In *Proceedings of the workshop on processing language variation: Digital armenian (DigitAm) within the 13th language resources and evaluation conference*, pages 38–42, Marseille, France. European Language Resources Association.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. 2022. Xtreme-s: Evaluating cross-lingual speech representations.

Hossep Dolatian, Afsheen Sharifzadeh, and Bert Vaux. 2023. *A grammar of Iranian Armenian: Parskahayeren or Iranahayeren*. Languages of the Caucasus. Language Science Press, Berlin. Unpublished manuscript.

Hossep Dolatian, Daniel Swanson, and Jonathan Washington. 2022. A Free/Open-Source morphological transducer for Western Armenian. In *Proceedings of the workshop on processing language variation: Digital armenian (DigitAm) within the 13th language resources and evaluation conference*, pages 1–7, Marseille, France. European Language Resources Association.

Narineh Hacopian. 2003. A three-way VOT contrast in final position: Data from Armenian. *Journal of the International Phonetic Association*, 33(1):51–80.

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020. Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the sixth international workshop on computational linguistics of Uralic languages*, pages 31–37, Wien, Austria. Association for Computational Linguistics.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada. Association for Computational Linguistics.

Niamh E. Kelly and Lara Keshishian. 2021. Voicing patterns in stops among heritage speakers of Western Armenian in Lebanon and the US. *Nordic Journal of Linguistics*, 44(2):103–129.

Karine Megerdoomian. 2009. Low-density language strategies for Persian and Armenian. In Sergei Nirenburg, editor, *Language Engineering for Lesser-Studied Languages*, pages 291–312. IOS Press, Amsterdam.

Jama Hussein Mohamud, Lloyd Acquaye Thompson, Aissatou Ndoye, and Laurent Besacier. 2021. Fast development of ASR in African languages using self supervised speech representation learning.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Avedis K Sanjian. 1996. The Armenian alphabet. In Peter T. Daniels and William Bright, editors, *The world's writing systems*, pages 356–363. Oxford University Press, New York and Oxford.

Scott Seyfarth, Hossep Dolatian, Peter Guekguezian, Niamh Kelly, and Tabita Toparlak. 2023. Armenian (Yerevan Eastern and Beirut Western varieties). *Journal of the International Phonetic Association*.

Scott Seyfarth and Marc Garellek. 2018. Plosive voicing acoustics and voice quality in Yerevan Armenian. *Journal of Phonetics*, 71:425–450.

Bert Vaux. 1998. *The phonology of Armenian*. Clarendon Press, Oxford.

Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 2020. Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th workshop on NLP for similar languages, varieties and dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

6. Language Resource References

Hossep Dolatian. 2024. *ReRooted: Speech corpus of Syrian Armenian refugee testimonials*. GitHub repository.

Khurshudian, Victoria G. and Daniel, Misha A. and Levonian, Dmitri V. and Plungian, Vladimir A. and Polyakov„ Alex E. and Rubakov, Sergey A. 2009. *Eastern Armenian National Corpus*. RGGU.

Khurshudyan, Victoria and Arkhangelskiy, Timofey and Daniel, Misha and Plungian, Vladimir and Levonian, Dmitri and Polyakov, Alex and Rubakov, Sergei. 2022. *Eastern Armenian national corpus: State of the art and perspectives*. European Language Resources Association.