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, [2,5]INALCO/SEDYL/CNRS , [4]Stony Brook University, INALCO

Introduction

Armenian is a language from **Indo-European family** and is from **independent branch** of that language family.

- Unique writing system
- 2 main dialects Eastern and Western Armenian
  - More than 10 sub-dialects (Yerevan, Lebanon, USA, Artsakh)
  - Eastern Armenian is the official language in the Republic of Armenia (Russia, Georgia, Iran)
  - Western Armenian developed in Ottoman Empire, became a diasporic dialect after the Armenian Genocide
- Nearly 7 million people speaks Armenian

Աա	Բբ	Գգ	Դդ	Եե	Զզ	Էէ	Ըը	Թթ	Ժժ
a	b	g	d	je	z	e	ë	t′	ž
Իի	Լլ	Խխ	Ծծ	Կկ	Հհ	Ձձ	Ղղ	Ճճ	Մմ
i	l	x	ts	k	h	dz	ŕ	tš	m
Յյ	Նն	Շշ	Ոո	Չչ	Պպ	Ջջ	Ռռ	Սս	Վվ
j	n	š	vo	tš′	p	dž	r	s	v
	Տտ	Րր	Ցց	Ււ	Փփ	Քք	Օօ	Ֆֆ	
	t	r′	ts′	w	p′	k′	o	f	

Objectives

1. Collect oral speech corpora for Eastern and Western Armenian
2. Develop ASR system for Eastern and Western Armenian
3. Research the abilities of joint ASR model that works both for Western and Eastern Armenian

Dialect Differences

- Pronunciation
  - Գ (g) - Ք (kʰ) - Կ (k)
  - Բ (b) - Պ (p) - Փ (pʰ)
  - Ձ (dz) - Ծ (ts) - Ց (tsʰ)
  - Դ (d) - Տ (t) - Թ (tʰ)
- Vocabulary
  - Turkish words (WA)
  - Russian words (EA)
- Accent

	Eastern	Western			
		Turkey	Lebanon	USA	
<բ> <բան>	**b**an	**p**ʰan	**p**an	**p**ʰan	‘thing’
<պ> <պահ>	**p**ah	**b**ah	**b**ah	**p**ah	‘period’
<փ> <փայլ>	**p**ʰajl	**p**ʰajl	**p**ajl	**p**ʰajl	‘shine’

Armenian Datasets

1. Common Voice
   a. Read speech data collected by volunteers
2. Google Fleurs
   a. Read speech where each sentence was pronounced by 3 native speakers
3. **EANC-based corpus** New!
4. **ReRooted** New!

Eastern Armenian National Corpus

- Contains 110 mln Eastern Armenian words
- Oral subcorpora contains 350 hrs
  a. Spontaneous Dialogs
  b. Task-oriented Narratives
  c. TV-shows
  d. Movies
- The alignment was done semi-automated with forced-alignment tool

ReRooted

- Western Armenian oral corpus of Syrian Armenian refugees
- Corpus contains 75 hrs of data
- Time-aligned caption
- The alignment errors were corrected manually with Praat TextGrid

Utilized Armenian 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.

Armenian Datasets at Present

Code	Source	Dialect	Speech type	Train	Dev	Test
CV	Common Voice	Eastern	Read	8,9 hr. (+3,4)	6,3 hr. (+2,3)	6,7 hr. (+2,2)
GF	Google Fleurs	Eastern	Read	10,5 hr.	1,2 hr.	3 hr.
EA	EANC	Eastern	Naturalistic	20,1 hr. (+14,3)	0,5 hr.	0,5 hr.
WA	ReRooted	Western	Naturalistic	10 hr. (+4,2)	0,5 hr.	0,5 hr.

Existing Armenian ASR Tools

1. ArmSpeech
2. Ican24
3. Arampacha
4. Whisper (Out-of-the-box)
5. Seamless M4T (Out-of-the-box)

Non of these tools support different Armenian dialects

Tested Models

1. Whisper-large-v1
2. Whisper-large-v2
3. Whisper-large-v3
4. SeamlessM4T-v1
5. SeamlessM4T-v2

Experiments

Type of data

- Eastern only vs. Western only vs. Bi-dialectal
- Naturalistic speech vs. Read speech vs. Both

Type of training

- Mimic that we have open source data pre-trained model
- Tuning from the checkpoint of base model

Experiments Scheme

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

→ denotes Fine-Tuning

Overall Results Analysis

1. Incorporating a specific dataset within the training set leads to an improvement in metrics for the corresponding test sets
2. More data the better model no matter the type of data
3. Open-source data trained models perform poorly on other domains

Language-transferability Results Analysis

1. Language-transferability is poor (WA-trained models performed poorly on EA data and vice versa)
2. The results were improved compared to Out-of-the-box => There is some knowledge transferability across dialects
3. EA and WA datasets utilized together achieve higher results
4. To achieve high result the specific dialect data is needed

Best Results

	Model	Training Scenario	WER	CER
Best WA model	Whisper-large-v2	CV + GF + EA + WA	**33,8**	**16,0**
Best OOB WA model	Seamless v1	-	76,6	44,5
Best EA model	Seamless v2	CV + GF ->EA	**29,0**	**18,9**
Best OOB EA model	Seamless v1	-	39,1	25,2
Best CV model	Seamless v2	CV + GF	**7,6**	**1,8**
Best GF model			**7,4**	**2,4**
Best OOB CV model	Seamless v2	-	10,5	3,2
Best OOB GF model			11,2	5,3
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**
Best OOB EA and WA Avg. model	Seamless v2	-	38,4	21,7
Best OOB all tests Avg. model			57,9	34,8

Best Results of Each Model (Avg.)

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

Error Analysis

1. Seamless misspells EA as WA **(a) (b)**
2. Whisper hearing WA transcribes it to non-existing word **(c)**
3. Whisper abbreviates some words **(d)**
4. Whisper omits words **(e)**

	Model	Audio (IPA)	Correct transcription	Model's incorrect transcription	Pronunciation of incorrect transcription
(a)	Seamless v2	/amarva/	ամառվա	ամառուայ	/amarva/
(b)	Seamless v2	/tʰalanvet͡sʰ/	թալանվեց	թալանուեցաւ	/tʰalanvet͡sʰav/
(c)	Whisper v3	/ɑrer ejiŋkʰ/	առեր էինք	առայինք	/arajiŋkʰ/
(d)	Whisper v3	/kilogramov/	կիլոգրամով	կգով	/kilogramov/
(e)	Whisper v3	/t͡sʰənt͡sʰumə mɛʁramisi pʰuln aveli/	ցնցումը մեղրամիսի փուլն ավելի	ցնցումը ավելի	t͡sʰənt͡sʰumə aveli

Future Work

1. Increasing amount of aligned data
2. Explore whether less amount of naturalistic speech data could achieve better results than more read speech data
3. Add more dialects

Thank you!