End-to-End Multilingual Automatic Speech Recognition for Less-Resourced Languages: The Case of Four Ethiopian Languages

Solomon Teferra Abate[1,2], Martha Yifiru Tachbelie[1,2] and Tanja Schultz[1]
[1]CSL, University of Bremen, Germany and [2]SIS, Addis Ababa University, Ethiopia
solomon.teferra,martha.yifiru@aau.edu.et
tanja.schultz@uni-bremen.de

Platzhalter QR-Code

Abstract

The End-to-End (E2E) approach, which maps a sequence of input features into a sequence of graphemes or words, to Automatic Speech Recognition (ASR) is a hot research agenda. It is interesting for less-resourced languages since it avoids the use of pronunciation dictionary, which is one of the major components in the traditional ASR systems. However, like any deep neural network (DNN) approaches, E2E is data greedy. This makes the application of E2E to less-resourced languages questionable. However, using data from other languages in a multilingual (ML) setup is being applied to solve the problem of data scarcity. We have, therefore, conducted ML E2E ASR experiments for four less-resourced Ethiopian languages using different language and acoustic modelling units. The results of our experiments show that relative Word Error Rate (WER) reductions (over the monolingual E2E systems) of up to 29.83\% can be achieved by just using data of two related languages in E2E ASR system training. Moreover, we have also noticed that the use of data from less related languages also leads to E2E ASR performance improvement over the use of monolingual data.

Introduction

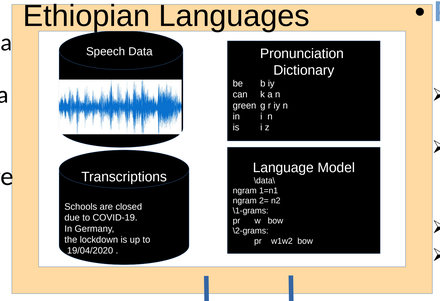
E2E Integrate LM

- Grapheme/ Words → Higher WER
- Grapheme/ Words → Lower WER

➔ Large Speech corpora
➔ Does not Exist in many languages
➢ The solution is application of MLASR

Monolingual E2E ASR

- **Writing System:**
  - Amharic and Tigrigna use Ethiopic Script
  - Oromo and Wolaytta use Latin
- **Morphology:**
  - All these languages are morphologically complex
    - Reflected in their high OOV rate



- **Phonology:**
  - Amharic and Tigrigna have 28 and 31 consonants, respectively
  - Amharic consonants are sub-sets of the Tigrigna phone set
  - Both share the same 7 vowels
  - Oromo and Wolaytta have 28 and 26 consonants, respectively
  - They share most of their consonants
  - They also share the same 5 vowels that have long and short forms

Char. E2E ASR | Phone E2E ASR

Languages	HMM-DNN	E2E ASR						
		Word LM		Char LM		Phone LM		
	WER	CER	WER	CER	WER	PER	WER
AMH2005	23.05	9.29	26.28	7.63	19.81	5.16	**19.05**
AMH2020	-	-	-	14.65	36.43	8.84	29.70
TIR	26.94	10.32	27.27	9.48	25.55	6.18	**22.18**
ORM	32.28	10.11	32.13	9.53	30.36	10.83	**30.36**
WAL	23.23	9.23	25.81	8.64	23.35	9.24	24.11

E2E MLASR

ML2: E2E MLASR

ML4: and ML23: E2E MLASR

Language/ Corpora	HMM-DNN	Character based		Phone based	
	WER	CER	WER	PER	WER
AMH2005	23.05	4.23	13.90	3.28	**13.63**
AMH2020	-	10.81	29.58	7.98	27.81
TIR	26.94	8.21	23.00	5.30	**20.91**
ORM	32.28	9.41	29.12	10.13	**29.01**
WAL	23.23	8.3	22.66	8.7	**21.04**

Language/ Corpora	Mono		ML4		ML23	
	PER	WER	PER	WER	PER	WER
AMH2005	5.16	**19.05**	3.26	14.06	3.12	**13.44**
AMH2020	8.84	29.70	7.26	26.16	7.67	29.34
TIR	6.18	**22.18**	5.12	21.14	5.49	22.43
ORM	10.83	**30.36**	8.7	27.37	9.17	28.86
WAL	9.24	24.11	6.5	**18.51**	6.63	19.91

Conclusions and Future Directions

- E2E is promising for the development of ASR systems for morphologically complex languages that suffer from very high OOV rates
- Characters and phones are generally better modeling units than small vocabulary words for language modeling
- E2E MLASR using resources related and even less related languages resulted in performance improvement over monolingual E2E systems
- Phone modeling units are better than characters for acoustic modeling

We propose:

- To extending this research for more target languages and
- To conduct experiments using E2E approach at different levels of data scarcity

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre