Dual Script E2E framework for Multilingual and Code-Switching ASR

MUCS 2021 Workshop

Authors: Mari Ganesh Kumar, Jom Kuriakose, Anand Thyagachandran, Arun Kumar A, Ashish Seth, Lodagala VSV Durga Prasad, Saish Jaiswal, Anusha Prakash and Hema Murthy

August 11, 2021

Presented by: Mari Ganesh Kumar

Outline

Introduction

Objective

To build better multilingual and code-switching ASR systems for low resource Indian languages

Multilingual ASR

```
Utt 1:      மேற்கு இந்தியத் தீவுகள் கொடி கட்டிப் பறந்த காலம் அது
Utt 2:      આવી જ ભાવનાત્મક વાતો રાહુલ ગાંધી પણ કરી રહ્યા છે
Utt 3:      सुशीला ने विमानचालकों को बताया कि उड़ान भरते हुए विमान कैसे गोता खाएँ
```

code-switching ASR

```
Utt 1:      libreoffice impress এর উপর এই কথ্য tutorial এ আপনাদের স্বাগত
Utt 2:      जैसे कि आप देख सकते हैं workspace में 5 tabs हैं जिन्हें view buttons कहते हैं
```

Approach

- Unified Parser for Indian languages (Baby et al. 2016a) - an in-house rule-based phoneme-level common label set (CLS) representation

Why CLS?

Approach

- Unified Parser for Indian languages (Baby et al. 2016a) - an in-house rule-based phoneme-level common label set (CLS) representation
- Use CLS representation to train multilingual and code-switching ASR

Why CLS?

Approach

- Unified Parser for Indian languages (Baby et al. 2016a) - an in-house rule-based phoneme-level common label set (CLS) representation
- Use CLS representation to train multilingual and code-switching ASR

Why CLS?

- (Prakash et al. 2019; Prakash and Murthy 2020) - Better text-to-speech (TTS) synthesis for low resource Indian languages using CLS

Approach

- Unified Parser for Indian languages (Baby et al. 2016a) - an in-house rule-based phoneme-level common label set (CLS) representation
- Use CLS representation to train multilingual and code-switching ASR

Why CLS?

- (Prakash et al. 2019; Prakash and Murthy 2020) - Better text-to-speech (TTS) synthesis for low resource Indian languages using CLS
- (Datta et al. 2020; Thomas, Audhkhasi, and Kingsbury 2020) - Use transliterated text to train multilingual ASR.

Approach

- Unified Parser for Indian languages (Baby et al. 2016a) - an in-house rule-based phoneme-level common label set (CLS) representation
- Use CLS representation to train multilingual and code-switching ASR

Why CLS?

- (Prakash et al. 2019; Prakash and Murthy 2020) - Better text-to-speech (TTS) synthesis for low resource Indian languages using CLS
- (Datta et al. 2020; Thomas, Audhkhasi, and Kingsbury 2020) - Use transliterated text to train multilingual ASR.
- (Shetty and Umesh 2021) - Use a character mapping between different Indian Languages, inspired by CLS, to train multilingual ASR

Common Label Set (CLS)

CLS

Table 1: Examples of words and their corresponding CLS representations

Language	Word	Parser output	CLS
Gujarati	હર્ષે	harsxee	harषE
Hindi	कड़वे	kadxwee	kaड़wE
Marathi	घट	ghatx	घaट
Odiya	ସାରିଛି	saarichi	sAriCi
Tamil	அணுமதி	anxumati	aणumati
Telugu	ఏంటీ	eeqtxii	Eqटी
English	action	AEKSHAHN	अkशan

Parsers

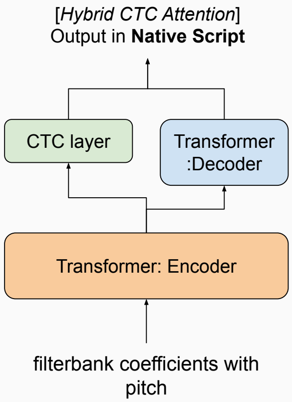
- Indian Languages: Unified Parser (Baby et al. 2016b)
- English : Neural network-based grapheme to phoneme converter (Park and Kim 2019)

Multilingual ASR

Baseline E2E Model

- Sampling Rate: 8000Hz
- Features: 80 mel filter bank energies along $+$ pitch
- Architectures: Hybrid CTC-attention models (Watanabe et al. 2018; Watanabe et al. 2017) using transformers
- Toolkit : ESPNet (Watanabe et al. 2018)
- Output Units: byte-pair (Kudo 2018) and character

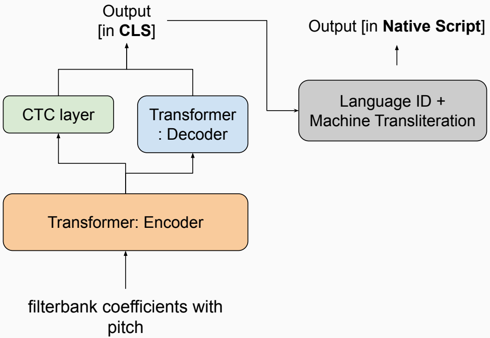
Table 2: Baseline Hybrid CTC-attention (Watanabe et al. 2018; Watanabe et al. 2017) model



CLS E2E Model

- CLS representation is used to pool data from all six languages and train E2E model

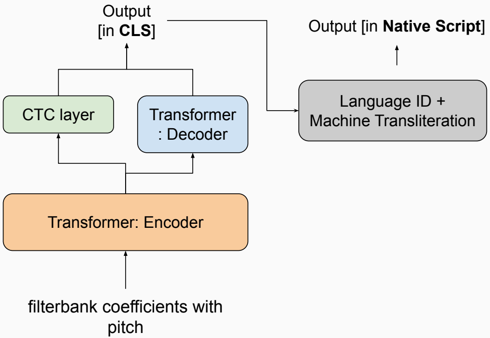
Table 3: Proposed CLS E2E model



CLS E2E Model

- CLS representation is used to pool data from all six languages and train E2E model
- Language ID is performed on the decoded CLS

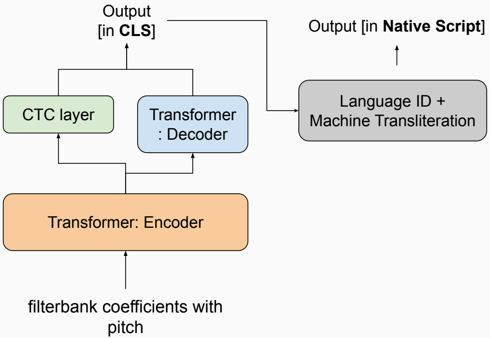
Table 3: Proposed CLS E2E model



CLS E2E Model

- CLS representation is used to pool data from all six languages and train E2E model
- Language ID is performed on the decoded CLS
- Machine Transliteration is used to retrieve native text from CLS

Table 3: Proposed CLS E2E model



Why Machine Transliteration?

- Phoneme (CLS) to grapheme (native script) mapping is not one-to-one
- Rules such as *schwa* deletion, geminate correction, and syllable parsing (Baby et al. 2016a) add to the complexity.

Table 4: Confusions in CLS to native script mapping

Language	CLS	Possible mappings
Hindi	kAmcor	कामचोर, काम्चोर
Bengali	sidধAnt	সিদ্ধান্ত, সিদধান্ত, সিদধানত

Language ID

- Features: multi-gram TF-IDF (at both character and word-level)
- Classifier: Naive Bayes
- Results: Accuracy of 99.7% on sub-task 1 development data
- Additional resources used: IndicTTS text data (Baby et al. 2016b)

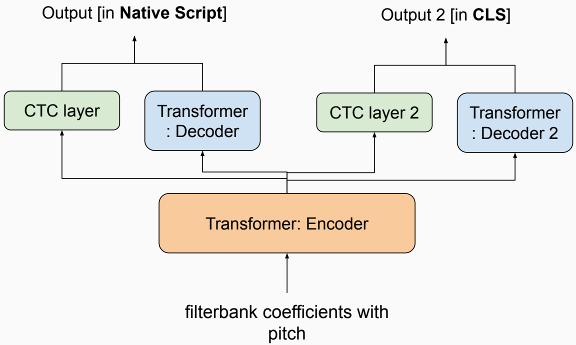
Machine Transliteration

- Toolkit: ONMT toolkit (Klein et al. 2017)
- Architecture: long short term memory (LSTM) based encoder-decoder model with global attention
- Results: 1.78% average WER and 0.44% average CER on sub-task 1 development data
- Additional resources used: IndicTTS text data (Baby et al. 2016b)

Dual Script E2E Model

- Integrates the LID and machine transliteration backend within the E2E model
- Two CTC layers and decoders to predict the CLS and native language script simultaneously

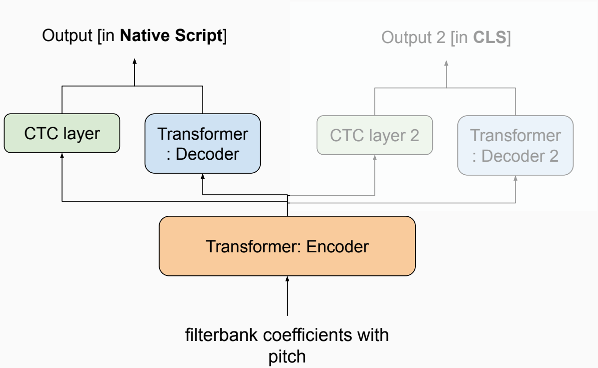
Figure 1: Proposed Dual Script E2E Model



Dual Script E2E Model

- Integrates the LID and machine transliteration backend within the E2E model
- Two CTC layers and decoders to predict the CLS and native language script simultaneously
- Output in CLS is discarded during decoding

Figure 1: Proposed Dual Script E2E Model



Results (Dual Script E2E Model)

Table 5: Results of sub-task 1 on development data

System Type	BPU/ CU	hi	mr	or	ta	te	gu	Avg
Challenge Baseline								
GMM-HMM	-	69.0	33.2	55.7	48.8	47.2	28.3	46.8
TDNN	-	40.4	22.4	39.0	33.5	**30.6**	**19.2**	30.7
Our Results (Without Language Model)								
Baseline E2E Model	BPU	52.1	33.8	71.3	31.3	32.9	26.5	49.5
	CU	26.5	17.1	36.1	35.3	36.6	28.4	30.0
CLS E2E Model	BPU	34	21.8	50.1	31.7	31.5	26.5	32.6
	CU	26.2	17.4	39.5	37.8	37.2	30.1	34.6
Dual Script E2E Model	BPU	29.4	19.8	44.9	**30.5**	31.9	24.4	30.1
	CU	**25.9**	**17.1**	**37.4**	35.2	35.8	27.7	29.8

- Dual script model has given the best performance for 4 out of 6 languages

Results (Dual Script E2E Model)

Table 5: Results of sub-task 1 on development data

System Type	BPU/ CU	hi	mr	or	ta	te	gu	Avg
Challenge Baseline								
GMM-HMM	-	69.0	33.2	55.7	48.8	47.2	28.3	46.8
TDNN	-	40.4	22.4	39.0	33.5	30.6	19.2	**30.7**
Our Results (Without Language Model)								
Baseline E2E Model	BPU	52.1	33.8	71.3	31.3	32.9	26.5	49.5
	CU	26.5	17.1	36.1	35.3	36.6	28.4	30.0
CLS E2E Model	BPU	34	21.8	50.1	31.7	31.5	26.5	32.6
	CU	26.2	17.4	39.5	37.8	37.2	30.1	34.6
Dual Script E2E Model	BPU	29.4	19.8	44.9	30.5	31.9	24.4	30.1
	CU	25.9	17.1	37.4	35.2	35.8	27.7	**29.8**

- Dual script model has given the best performance for 4 out of 6 languages
- The CU Dual Script model improves the average WER by $\approx 1\%$ over baseline without using any language model

Language Model

- Architecture: Transformers
- Toolkit: ESPNet (Watanabe et al. 2018)
- Additional Resources Used: Indic TTS (Baby et al. 2016b), AI4Bharat NLP corpora (Kakwani et al. 2020)
- Total Size: ≈ 150 million sentences
- No of epochs trained: 1

Results with LM

Table 6: Results of sub-task 1 on development data

System Type	BPU/ CU	hi	mr	or	ta	te	gu	Avg
TDNN	-	40.4	22.4	39.0	33.5	30.6	19.2	30.7
Results without any Language Model								
CLS	BPU	34	21.8	50.1	31.7	31.5	26.5	32.6
E2E Model	CU	26.2	17.4	39.5	37.8	37.2	30.1	34.6
Dual Script	BPU	29.4	19.8	44.9	30.5	31.9	24.4	30.1
E2E Model	CU	25.9	17.1	37.4	35.2	35.8	27.7	29.8
Results with Language Model								
CLS	BPU	31.8	21.8	48.2	25.6	24.2	20.7	28.7
E2E Model	CU	**21.4**	**14.6**	38.3	28.8	27.3	22.4	25.4
Dual Script	BPU	27.8	20.0	48.2	**23.6**	**23.6**	**18.8**	27.0
E2E Model	CU	21.6	15.1	**36.0**	25.9	25.3	20.5	24.0

- With language model, the proposed model gave better results for all the six languages

Results with LM

Table 6: Results of sub-task 1 on development data

System Type	BPU/ CU	hi	mr	or	ta	te	gu	Avg
TDNN	-	40.4	22.4	39.0	33.5	30.6	19.2	**30.7**
Results without any Language Model								
CLS E2E Model	BPU	34	21.8	50.1	31.7	31.5	26.5	32.6
	CU	26.2	17.4	39.5	37.8	37.2	30.1	34.6
Dual Script E2E Model	BPU	29.4	19.8	44.9	30.5	31.9	24.4	30.1
	CU	25.9	17.1	37.4	35.2	35.8	27.7	29.8
Results with Language Model								
CLS E2E Model	BPU	31.8	21.8	48.2	25.6	24.2	20.7	28.7
	CU	21.4	14.6	38.3	28.8	27.3	22.4	25.4
Dual Script E2E Model	BPU	27.8	20.0	48.2	23.6	23.6	18.8	27.0
	CU	21.6	15.1	36.0	25.9	25.3	20.5	**24.0**

- With language model, the proposed model gave better results for all the six languages
- The CU Dual Script model achieved an absolute improvement of 6% over the challenge baseline.

Results with LM

Table 6: Results of sub-task 1 on development data

System Type	BPU/ CU	hi	mr	or	ta	te	gu	Avg
TDNN	-	40.4	22.4	39.0	33.5	30.6	19.2	**30.7**
Results without any Language Model								
CLS E2E Model	BPU	34	21.8	50.1	31.7	31.5	26.5	32.6
	CU	26.2	17.4	39.5	37.8	37.2	30.1	34.6
Dual Script E2E Model	BPU	29.4	19.8	44.9	30.5	31.9	24.4	30.1
	CU	25.9	17.1	37.4	35.2	35.8	27.7	29.8
Results with Language Model								
CLS E2E Model	BPU	31.8	21.8	48.2	25.6	24.2	20.7	28.7
	CU	21.4	14.6	38.3	28.8	27.3	22.4	25.4
Dual Script E2E Model	BPU	27.8	20.0	48.2	23.6	23.6	18.8	27.0
	CU	21.6	15.1	36.0	25.9	25.3	20.5	**24.0**

- With language model, the proposed model gave better results for all the six languages
- The CU Dual Script model achieved an absolute improvement of 6% over the challenge baseline.
- The best performing three systems were submitted for evaluation on blind data

Blind Test Results

Table 7: Results of sub-task 1 on blind data

System Type	BPU /CU	hi	mr	or	ta	te	gu	Avg
Challenge Baseline								
TDNN	-	37.2	**29.0**	38.4	34.0	31.4	26.1	**32.73**
Submitted Systems								
CLS E2E Model	CU	19.5	85.9	37.1	32.0	30.3	32.9	39.6
Dual Script E2E Model	BPU	25.3	100.3	51.2	**25.1**	**25.4**	**25.4**	42.1
Dual Script E2E Model	CU	**17.8**	111.7	**32.1**	27.1	28.1	29.8	41.1

- Except Marthi, the dual script system outperformed the baseline results for all six languages
- On the average WER, still the baseline was better

Blind Test Results without Marathi

Table 8: Results of sub-task 1 on blind data

System Type	BPU/ CU	hi	or	ta	te	gu	Avg
Challenge Baseline							
TDNN	-	37.2	38.4	34.0	31.4	26.1	33.4
Submitted Systems							
CLS E2E Model	CU	19.5	37.1	32.0	30.3	32.9	30.3
Dual Script E2E Model	BPU	25.3	51.2	**25.1**	**25.4**	**25.4**	30.4
Dual Script E2E Model	CU	**17.8**	**32.1**	27.1	28.1	29.8	27.4

- Excluding Marathi, the submitted system achieved 6% absolute improvement in average WER.

Code Switching ASR

Sub Task 2 Results

Table 9: Results of sub-task 2 on development and blind data

System Type	BPU/ CU	Dev Data		
		hi-en	bn-en	Avg
Challenge Baseline (With Language Model)				
GMM-HMM	-	44.3	39.1	41.7
TDNN	-	36.9	34.0	35.6
E2E Model	BPU	27.7	37.2	32.4
Our Results (Without Language Model)				
Dual Script E2E Model	CU	33.0	27.0	30
	BPU	28.9	**25.3**	**27.1**

- For subtask 2, BPU give better results consistently
- The proposed model achieved 5% improvement over the challenge baseline without any language model
- Note: Baseline systems were trained and decoded separately for each language-pair. The proposed system were trained combinedly and language-pair information was not given during decoding.

Sub Task 2 Blind Test Result

Table 10: Results of sub-task 2 on development and blind data

System Type	BPU/ CU	Blind Test		
		hi-en	bn-en	Avg
Challenge Baseline (With Language Model)				
E2E Model	BPU	25.5	32.8	29.1
Our Results (Without Language Model)				
Dual Script E2E Model	BPU	**22.0**	27.8	24.9

- On the blind test as well, the proposed model achieved 5% improvement over the challenge baseline.

Other Issues

Figure 2: Example of a wrongfully penalized utterance

REF: கிருபா நான் நலம் உங்கள் நலம் அறிய ஆவல் இப்படிக்கு
உங்கள் *************** உன்மையுள்ள காதலி
HYP: திருபா நான் நலம் உங்கள் ************ நலமறிய ஆவல் இப்படிக்கு
உங்கள் உண்மை உள்ள காதலி

Figure 3: Sentences mostly made up of English words (From subtask-1 Marathi dataset)

REF: एचडीएफसी बँक शेअर अनॅलिसिस शो कर
(HDFC Bank Share Analysis Show)

REF: सबस्क्राईब रिन्यू कर
(Subscribe Renew)

REF: जिओ अॅप डाऊनलोड कर
(JIO app download)

Valid Languages: Hindi, Marathi, Gujarati

Summary

Summary

- Using two different models, CLS representation has been shown to be effective for both multilingual and code-switching task in the context of ASR
- Dual Script framework provides a novel way to train multilingual ASR using the native script and a common representation

References

- Baby, Arun et al. (2016a). “A unified parser for developing Indian language text to speech synthesizers”. In: *International Conference on Text, Speech and Dialogue*, pp. 514–521.
- Baby, Arun et al. (2016b). “Resources for Indian languages”. In: *Community-based Building of Language Resources (International Conference on Text, Speech and Dialogue)*, pp. 37–43.
- Datta, Arindrima et al. (2020). “Language-agnostic multilingual modeling”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8239–8243.
- Diwan, Anuj et al. (2021). “Multilingual and code-switching ASR challenges for low resource Indian languages”. In: *arXiv preprint arXiv:2104.00235*.

- Kakwani, Divyanshu et al. (2020). "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages". In: *Findings of EMNLP*.
- Klein, Guillaume et al. (2017). "Opennmt: Open-source toolkit for neural machine translation". In: *arXiv preprint arXiv:1701.02810*.
- Kudo, Taku (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: arXiv: `1804.10959 [cs.CL]`.
- Park, Kyubyong and Jongseok Kim (2019). *g2pE*. `https://github.com/Kyubyong/g2p`.
- Prakash, Anusha and Hema A. Murthy (2020). "Generic Indic Text-to-Speech Synthesisers with Rapid Adaptation in an End-to-End Framework". In: *Interspeech*, pp. 2962–2966.
- Prakash, Anusha et al. (2019). "Building Multilingual End-to-End Speech Synthesisers for Indian Languages". In: *10th ISCA Speech Synthesis Workshop (SSW)*, pp. 194–199.
- Shetty, Vishwas M and S Umesh (2021). "Exploring the use of Common Label Set to Improve Speech Recognition of Low Resource Indian Languages". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7228–7232.

Thomas, Samuel, Kartik Audhkhasi, and Brian Kingsbury (2020). "Transliteration Based Data Augmentation for Training Multilingual ASR Acoustic Models in Low Resource Settings". In: *Proc. Interspeech 2020*, pp. 4736–4740.

Watanabe, Shinji et al. (2017). "Hybrid CTC/attention architecture for end-to-end speech recognition". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8, pp. 1240–1253.

Watanabe, Shinji et al. (2018). "ESPnet: End-to-End Speech Processing Toolkit". In: *Proceedings of Interspeech*, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456. URL: http://dx.doi.org/10.21437/Interspeech.2018-1456.

Thank You

&

Questions?