### REED: An Approach Towards Quickly Bootstrapping Multilingual Acoustic Models

Bipasha Sen<sup>1\*,</sup> Aditya Agarwal<sup>1\*</sup>, Mirishkar Sai Ganesh<sup>2</sup>, Anil Kumar Vuppala<sup>2</sup>

Microsoft India Development Center, Hyderabad<sup>[1]</sup>

International Institute of Information Technology, Hyderabad<sup>[2]</sup> bise@microsoft.com, adiagar@microsoft.com, mirishkar.ganesh@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Spoken Language Technology 2021





INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

# Joint Acoustic Modeling

- Language Agnostic Layers.
- Current approaches –

SOTA  $\rightarrow$  complex architectures like RNN, Transformers.

### <u>Challenge</u> – Measuring combability of languages.

Ex. Tamil, Multilingual(Tamil, Telugu) > Multilingual (Tamil, Telugu, Gujarati)\*.

#### Automating compatibility -

- 1. Dependency on linguistic experts.
- 2. Running multiple combinations of languages on SOTA architectures (inefficient).

### <u>Solution</u> – REED – a fast joint multilingual acoustic model.

# rS. Joint Acoustic Model

# REED: Based on 1D Convolutional Layers.

#### **Computationally Efficient:**

- Based on simple 1D convolutional layers (feature extraction layer)
- 2. Exploits very short-term context unlike longterm context in SOTA architectures based on sequential models\*.

Ex. – cat (k-**ae**-t), car (k-**aa**-r), hack (hh-**ae**-k), sky (s-k-**ay**); "this is a cat" (dh-ih-s ih-z**ah** k-**ae**-t)

#### **Improved Accuracy:**

Operates directly on raw speech signals -

- 1. Improves accuracy
- 2. Learns suitable features instead of using perceptual handcrafted features.



## Experimental Results -- Accuracy

- Greater context on the left produces better WER compared to greater context on the right indicating that the left context contains more relevant information compared to the right.
- Even with full context (5 + 5), REED on MFCC performs at par with the most basic REED on raw speech signals with no context (0 + 0). This validates our hypothesis that CNNs can capture richer representation of the raw speech signals than traditional handcrafted MFCC features

_	Models + Context	Gujarati	Telugu	Tamil
_	lstm + mfcc	16.12	20.24	19.86
1	$cnn+raw + \{0,0\}$	24.06	31.23	30.94
$\sqrt{-}$	cnn+raw + $\{-1, +1\}$	23.92	30.66	29.59
$\langle \rangle$	cnn+raw + $\{-2, +1\}$	20.13	26.80	25.73
	cnn+raw + $\{-1, +2\}$	22.65	27.32	26.93
_	cnn+raw + $\{-2, +2\}$	19.48	22.37	21.55
	cnn+raw + $\{-3, +2\}$	18.36	21.24	20.92
_	cnn+raw + $\{-2, +3\}$	19.02	22.33	20.98
*	cnn+mfcc + $\{-5, +5\}$	25.05	31.13	30.78

WER of different architectural configurations in %.

REED operated on raw speech signals.

The number inside the braces indicate the number of left and right context windows, respectively.

 REED operated on MFCC features extracted on Speech utterances.

# Experimental Results – Computational Boost

Computational Boost and Accuracy tradeoff -- An interesting trend of the decrease in training and inference time with the decrease in the context window is observed. Increase in context size improved the accuracy.

				Train	ing	Infe	erence
Exp	Models + Context	Avg. WER	Avg. WER deg.	Time	Speed Up	Time	Speed Up
1	lstm + mfcc	18.74	-	$\sim$ 4.5 days	-	780 ms	-
2	$cnn+raw + \{0,0\}$	28.74	-10	↑ 7.84 hours	$\sim 13.5 \times$	15ms	$\sim 52 \times$
3	cnn+raw + $\{-1, +1\}$	28.05	-9.31	11.56 hours	$\sim 9  imes$	29ms	$\sim 26 \times$
4	cnn+raw + $\{-2, +1\}$	24.22	-5.48	19.08 hours	$\sim 5.6 \times$	30ms	$\sim 26 \times$
5	cnn+raw + $\{-1, +2\}$	25.63	-6.89	20.63 hours	$\sim 5  imes$	32ms	$\sim 24 \times$
6	cnn+raw + $\{-2, +2\}$	21.13	-2.39	24.50 hours	$\sim 4.4 \times$	89ms	$\sim 8.7  imes$
7	cnn+raw + $\{-3, +2\}$	20.17	-1.43	27.29 hours	$\sim 4 \times$	105ms	$\sim 7.4 \times$
8	cnn+raw + $\{-2, +3\}$	• 20.77	-2.03	28.62 hours	$\sim 3.5 \times$	109ms	$\sim 7.15 \times$
9	cnn+mfcc + $\{-5, +5\}$	28.98	-10.24	3.55 hours	$\sim 30  imes$	12 ms	$\sim 65 \times$

The number inside the braces indicate the number of left and right context windows, respectively.

# Experimental Results – Computational Boost

• Best average WER -- 20.17 on REED with raw speech signals, 3 left - 2 right context.

7	cnn+raw + $\{-3, +2\}$	20.17	-1.43	27.29 hours	$\sim 4 \times$	105ms	$\sim 7.4 \times$
---	------------------------	-------	-------	-------------	-----------------	-------	-------------------

• Baseline

1	lstm + mfcc	18.74	-	$\sim$ 4.5 days	-	780 ms	-
---	-------------	-------	---	-----------------	---	--------	---

- Comparable WER to the baseline -- an average relative degradation of only <u>~7%</u>.
- Computation boost of <u>4x</u> in <u>training</u>: ~28 hours training time against ~4.5 days training time on baseline.
- Computational boost of <u>7.4x</u> in <u>inference</u>: an average of 105ms to infer on an average of 5.85s long speech utterance compared to 780ms on the baseline.
- There is a significant jump in the WER between Exp 4 and Exp 3 suggesting no-context or a context window of 1 on either side doesn't capture the relevant coarticulations.

2	$cnn+raw + \{0,0\}$	28.74	-10	7.84 hours	$\sim 13.5 \times$	15ms	$\sim 52 \times$
3	cnn+raw + $\{-1, +1\}$	28.05	-9.31	11.56 hours	$\sim 9 \times$	29ms	$\sim 26 \times$
4	cnn+raw + $\{-2, +1\}$	24.22	-5.48	19.08 hours	$\sim 5.6 \times$	30ms	$\sim 26 \times$

# Conclusion

- REED is a hybrid architecture based on a simple acoustic model which uses only <u>1D convolutional layers</u> and <u>feed-forward networks</u> and operates on <u>very short context</u>.
- REED is a <u>fast-multilingual system</u> that is an effective means for <u>quickly bootstrapping</u> and <u>validating</u> the <u>compatibility of different languages</u> for building a robust multilingual system.
- <u>CNNs</u> as a <u>feature extraction layer</u> can be used to learn <u>rich representation</u> of the raw speech signals instead of relying on traditional hand-crafted features like MFCCs.
- Most optimal model (with 3 left and 2 right context) provided a training and inference boost of 4x and 7.4x respectively with a relative degradation of only ~7% against the baseline.
- Depending on the <u>requirement</u> of the system and <u>acceptable WER degradation</u>, one of <u>REED's configuration</u> can be employed to build <u>robust</u> multilingual systems without additional <u>linguistic knowledge</u>.

## This is the end of my highlight

### REED: An Approach Towards Quickly Bootstrapping Multilingual Acoustic Models

Bipasha Sen<sup>1\*,</sup> Aditya Agarwal<sup>1\*</sup>, Mirishkar Sai Ganesh<sup>2</sup>, Anil Kumar Vuppala<sup>2</sup>

Microsoft India Development Center, Hyderabad<sup>[1]</sup>

International Institute of Information Technology, Hyderabad<sup>[2]</sup> bise@microsoft.com, adiagar@microsoft.com, mirishkar.ganesh@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Spoken Language Technology 2021





INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

# What are multilingual systems?

• A single system capable of transcribing speech for multiple languages.

### • Use Cases –

- 1. Low resource languages
- 2. Code switched languages Example -- Hinglish
- Existing approaches
  - 1. Shared hidden layers
  - 2. Bottleneck layers
  - 3. Multitask learning
  - 4. Joint acoustic modelling.



Bottleneck Layers

# Existing Approaches -- Drawbacks

 Language dependent layers – Shared Hidden Layers / Multitask Learning / Bottleneck Layers.



# Joint Acoustic Modeling

- Language Agnostic Layers.
- Current approaches –

SOTA  $\rightarrow$  complex architectures like RNN, Transformers.

### <u>Challenge</u> – Measuring combability of languages.

Ex. Tamil, Multilingual(Tamil, Telugu) > Multilingual (Tamil, Telugu, Gujarati)\*.

#### Automating compatibility -

- 1. Dependency on linguistic experts.
- 2. Running multiple combinations of languages on SOTA architectures (inefficient).

### <u>Solution</u> – REED – a fast joint multilingual acoustic model.

# rS. Joint Acoustic Model

# REED: Based on 1D Convolutional Layers.

#### **Computationally Efficient:**

- Based on simple 1D convolutional layers (feature extraction layer)
- 2. Exploits very short-term context unlike longterm context in SOTA architectures based on sequential models\*.

Ex. – cat (k-**ae**-t), car (k-**aa**-r), hack (hh-**ae**-k), sky (s-k-**ay**); "this is a cat" (dh-ih-s ih-z**ah** k-**ae**-t)

#### **Improved Accuracy:**

Operates directly on raw speech signals -

- 1. Improves accuracy
- 2. Learns suitable features instead of using perceptual handcrafted features.



### Dataset

SpeechOcean's low resource automatic speech recognition for Indian languages\*.

- 1. Three Indic languages -- Gujarati, Telugu, and Tamil spoken by multiple speakers.
- 2. 120 hours of spoken utterances: 40 hours of each languages.
- 3. Test and validation data: 5 hours per language.

Languages	Train Set	Dev Set	Test Set
Gujarati	18307	4500	3075
Telugu	41682	3200	3040
Tamil	35231	3900	3081

Number of utterances per languages

\*Brij Srivastava, Sunayana Sitaram, Rupesh Mehta, Krishna Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for Indian languages"

## Experimental Results -- Accuracy

- Greater context on the left produces better WER compared to greater context on the right indicating that the left context contains more relevant information compared to the right.
- 2. Even with full context (5 + 5), REED on MFCC performs at par with the most basic REED on raw speech signals with no context (0 + 0). This validates our hypothesis that CNNs can capture richer representation of the raw speech signals than traditional handcrafted MFCC features

_	Models + Context	Gujarati	Telugu	Tamil
_	lstm + mfcc	16.12	20.24	19.86
1	$cnn+raw + \{0,0\}$	24.06	31.23	30.94
	cnn+raw + $\{-1, +1\}$	23.92	30.66	29.59
	cnn+raw + $\{-2, +1\}$	20.13	26.80	25.73
	cnn+raw + $\{-1,+2\}$	22.65	27.32	26.93
	cnn+raw + $\{-2, +2\}$	19.48	22.37	21.55
	cnn+raw + $\{-3, +2\}$	18.36	21.24	20.92
_	cnn+raw + $\{-2, +3\}$	19.02	22.33	20.98
	cnn+mfcc + $\{-5, +5\}$	25.05	31.13	30.78

WER of different architectural configurations in %.

The number inside the braces indicate the number of left and right context windows, respectively.

REED operated on raw speech signals.

 REED operated on MFCC features extracted on Speech utterances.

# Experimental Results – Computational Boost

Computational Boost and Accuracy tradeoff -- An interesting trend of the decrease in training and inference time with the decrease in the context window is observed. Increase in context size improved the accuracy.

				Train	ing	Infe	erence
Exp	Models + Context	Avg. WER	Avg. WER deg.	Time	Speed Up	Time	Speed Up
1	lstm + mfcc	18.74	-	$\sim$ 4.5 days	-	780 ms	-
2	$cnn+raw + \{0,0\}$	28.74	-10	↑ 7.84 hours	$\sim 13.5 \times$	15ms	$\sim 52 \times$
3	cnn+raw + $\{-1, +1\}$	28.05	-9.31	11.56 hours	$\sim 9  imes$	29ms	$\sim 26 \times$
4	cnn+raw + $\{-2, +1\}$	24.22	-5.48	19.08 hours	$\sim 5.6 \times$	30ms	$\sim 26 \times$
5	cnn+raw + $\{-1, +2\}$	25.63	-6.89	20.63 hours	$\sim 5  imes$	32ms	$\sim 24 \times$
6	cnn+raw + $\{-2, +2\}$	21.13	-2.39	24.50 hours	$\sim 4.4 \times$	89ms	$\sim 8.7  imes$
7	cnn+raw + $\{-3, +2\}$	20.17	-1.43	27.29 hours	$\sim 4 \times$	105ms	$\sim 7.4 \times$
8	cnn+raw + $\{-2, +3\}$	• 20.77	-2.03	28.62 hours	$\sim 3.5 \times$	109ms	$\sim 7.15 \times$
9	cnn+mfcc + $\{-5, +5\}$	28.98	-10.24	3.55 hours	$\sim 30  imes$	12 ms	$\sim 65 \times$

The number inside the braces indicate the number of left and right context windows, respectively.

# Experimental Results – Computational Boost

• Best average WER -- 20.17 on REED with raw speech signals, 3 left - 2 right context.

7	cnn+raw + $\{-3, +2\}$	20.17	-1.43	27.29 hours	$\sim 4 \times$	105ms	$\sim 7.4 \times$
---	------------------------	-------	-------	-------------	-----------------	-------	-------------------

• Baseline

1	lstm + mfcc	18.74	-	$\sim$ 4.5 days	-	780 ms	-
---	-------------	-------	---	-----------------	---	--------	---

- Comparable WER to the baseline -- an average relative degradation of only <u>~7%</u>.
- Computation boost of <u>4x</u> in <u>training</u>: ~28 hours training time against ~4.5 days training time on baseline.
- Computational boost of <u>7.4x</u> in <u>inference</u>: an average of 105ms to infer on an average of 5.85s long speech utterance compared to 780ms on the baseline.
- There is a significant jump in the WER between Exp 4 and Exp 3 suggesting no-context or a context window of 1 on either side doesn't capture the relevant coarticulations.

2	$cnn+raw + \{0,0\}$	28.74	-10	7.84 hours	$\sim 13.5 \times$	15ms	$\sim 52 \times$
3	cnn+raw + $\{-1, +1\}$	28.05	-9.31	11.56 hours	$\sim 9 \times$	29ms	$\sim 26 \times$
4	cnn+raw + $\{-2, +1\}$	24.22	-5.48	19.08 hours	$\sim 5.6 \times$	30ms	$\sim 26 \times$

## Use Case -- Low Resource Languages

#### • Examples –

Indian languages (Hindi + Marathi + Gujarati + Tamil + Telugu); South African languages (Zulu + Xhosa + Sesotho).

- Indian context
  - 1500 languages in India
  - 22 languages accorded official status
  - 30 languages with more than 1 million speakers
  - Low resource Bengali, Marathi, Gujarati, Telugu etc.



- Building a robust monolingual acoustic system requires a <u>sizeable amount</u> of training data.
- Employed <u>**REED</u>** for quickly bootstrapping a multilingual system for low-resource languages.</u>

# Experimental Results – Linguistic Experiments

- Multilingual Acoustic Model for 3 Indic languages Gujarati, Tamil, and Telugu.
- Gujarati Indo-Aryan family; Tamil and Telugu Indo-Dravidian family.

#### **Experimental Setup** --

1. Built a monolingual system for each of the three languages and noted the individual monolingual WERs.

		Multilingual			
Languages Monolingual		ta + te	ta + te + gu		
Gujarati	19.88	-	18.36		
Telugu	29.07	20.85	21.24		
Tamil	28.91	20.67	20.92		

## Experimental Results – Linguistic Experiments

- Multilingual Acoustic Model for 3 Indic languages Gujarati, Tamil, and Telugu.
- Gujarati Indo-Aryan family; Tamil and Telugu Indo-Dravidian family.

#### **Experimental Setup** --

- 1. Built a monolingual system for each of the three languages and noted the individual monolingual WERs.
- Combined the two Dravidian languages Tamil and Telugu (Model ta + te).
  This increases the phoneme overlap across these two languages.

		Multilingual	
Languages	Monolingual	ta + te	ta + te + gu
Gujarati	19.88	-	18.36
Telugu	29.07	20.85	21.24
Tamil	28.91	20.67	20.92

# Experimental Results – Linguistic Experiments

- Multilingual Acoustic Model for 3 Indic languages Gujarati, Tamil, and Telugu.
- Gujarati Indo-Aryan family; Tamil and Telugu Indo-Dravidian family.

#### **Experimental Setup** --

- 1. Built a monolingual system for each of the three languages and noted the individual monolingual WERs.
- Combined the two Dravidian languages Tamil and Telugu (Model *ta + te*).
  This increases the phoneme overlap across these two languages.
- 3. Added Gujarati, that belongs to Indo-Aryan family, to the combined dataset of Tamil and Telugu (Model *ta* + *te* + *gu*)

		Multilingual	
Languages	Monolingual	ta + te	ta + te + gu
Gujarati	19.88	-	18.36
Telugu	29.07	20.85	21.24
Tamil	28.91	20.67	20.92

# Observations

		Multilingual	
Languages	Monolingual	ta + te	ta + te + gu
Gujarati	19.88	-	18.36
Telugu	29.07	20.85	21.24
Tamil	28.91	20.67	20.92

- This validates the hypothesis that using a combination of languages belonging to the <u>same language family</u> should give better WERs compared to a system built by combining the languages belonging to different language families (e.g., Indo-Aryan and Indo-Dravidian).
- The <u>performance</u> of a multilingual system is directly dependent on the <u>compatibility</u> of the languages.
- <u>Familial correspondence</u> maybe more important than <u>additional data</u> provided by the pooled languages.
- Even though combining languages with shared phone space can improve the accuracy of the individual languages, <u>smart</u> <u>selection of languages</u> used to build multilingual systems can further decrease the WERs for the individual languages.

# Conclusion

- REED is a hybrid architecture based on a simple acoustic model which uses only <u>1D convolutional layers</u> and <u>feed-forward networks</u> and operates on <u>very short context</u>.
- REED is a <u>fast-multilingual system</u> that is an effective means for <u>quickly bootstrapping</u> and <u>validating</u> the <u>compatibility of different languages</u> for building a robust multilingual system.
- <u>CNNs</u> as a <u>feature extraction layer</u> can be used to learn <u>rich representation</u> of the raw speech signals instead of relying on traditional hand-crafted features like MFCCs.
- Most optimal model (with 3 left and 2 right context) provided a training and inference boost of 4x and 7.4x respectively with a relative degradation of only ~7% against the baseline.
- Experimental results show that <u>familial correspondence</u> maybe more important than <u>additional data</u> provided by pooled languages.
- Depending on the <u>requirement</u> of the system and <u>acceptable WER degradation</u>, one of <u>REED's configuration</u> can be employed to build <u>robust</u> multilingual systems without additional <u>linguistic knowledge</u>.

## Acknowledgements

We would like to thank

- 1. TDIL MeitY for supporting this work through "Crowd Sourcing of Large Speech Data Sets To Enable Indian Language Speech - Speech Solutions (Pilot Project)".
- 2. Dr. Rajeev Gupta, Dr. Sandipan Dandapat, and Dr. Sunayana Sitaram, who are researchers at Microsoft, for their valuable feedback.

### Thanks! We are open to questions ③