End-to-End Language Recognition Using Attention Based Hierarchical Gated Recurrent Unit Models

Bharat Padi[1], **Anand Mohan**[2], **Sriram Ganapathy**[2]

[1]minds.ai, Bengaluru

[2]LEAP Lab, Department of Electrical Engineering, Indian Institute of Science, Bengaluru

Motivation

- Certain regions of the audio can be more important than the rest.
- Conventional approaches (i-vector and x-vector) ignore the sequence information.
- Previous end-to-end approaches work well only on short durations (3 sec) [1].

Proposed HGRU Model

- Hierarchically builds a sequence of 1 sec representations.
- Attention module computes a weighted average of this sequence to output utterance level embedding.
- Duration dependent fully connected layers compute posteriors from the embedding.

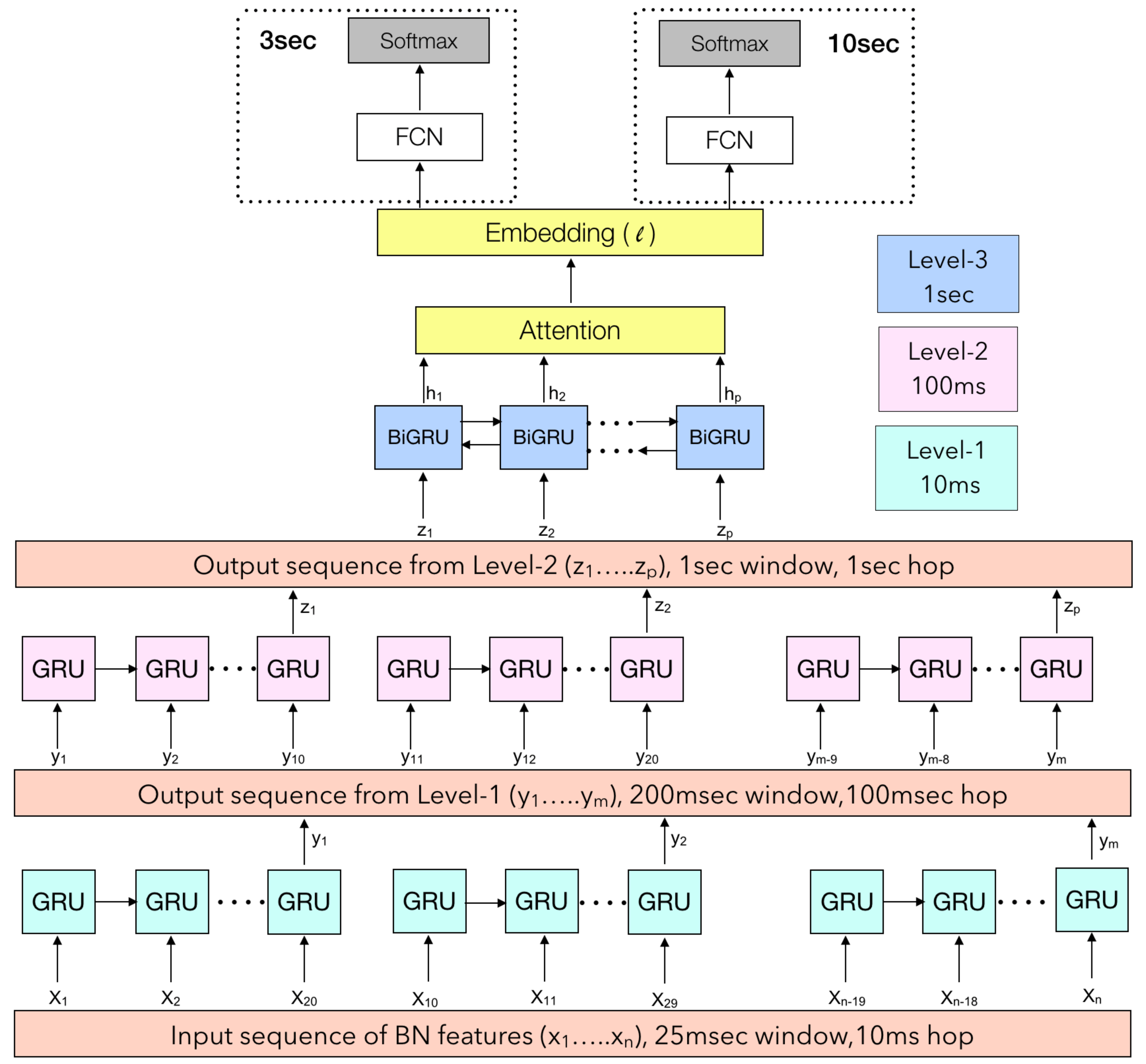


Figure 1: Proposed HGRU Model

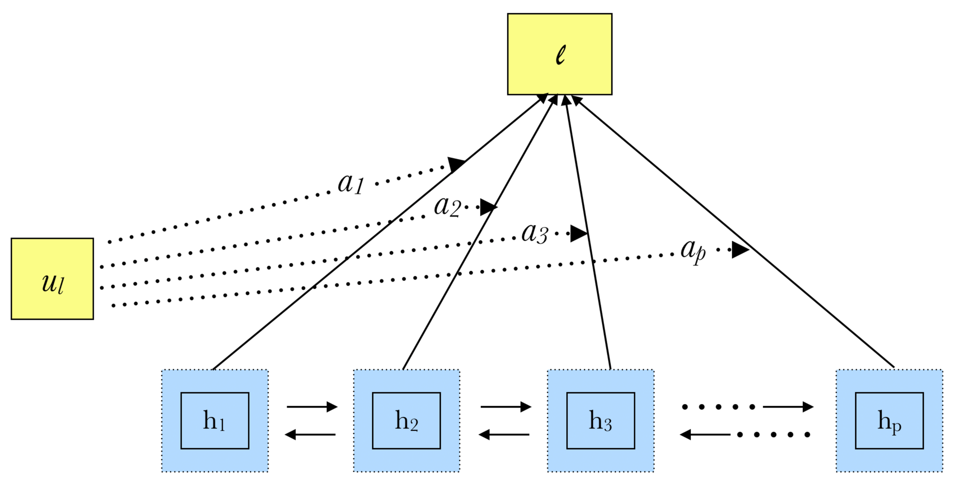


Figure 2: Attention Module

$$\mathbf{u}_t = tanh(\mathbf{W}_l\mathbf{h}_t + \mathbf{b}_l)$$

$$a_t = \frac{exp(\mathbf{u}_t^T\mathbf{u}_l)}{\sum_t exp(\mathbf{u}_t^T\mathbf{u}_l)}$$

$$\boldsymbol{l} = \sum_t a_t\mathbf{h}_t$$

Experiments

Cluster	Target Languages	Hours
Arabic	Egyptian Arabic (ara-arz)	190.9
	Iraqi Arabic (ara-acm)	130.8
	Levantine Arabic (ara-apc)	440.7
	Maghrebi Arabic (ara-ary)	81.8
Chinese	Mandarin (zho-cmn)	379.4
	Min Nan (zho-nan)	13.3
English	British English (eng-gbr)	4.8
	General American English (eng-usg)	327.7
Slavic	Polish (qsl-pol)	59.3
	Russian (qsl-rus)	69.5
Iberian	Caribbean Spanish (spa-car)	166.3
	European Spanish (spa-eur)	24.7
	Latin American Continental Spanish (spa-lac)	175.9
	Brazilian Portuguese (por-brz)	4.1

Table 1: LRE17 training set : target languages, language clusters and total number of hours.

- Experiments performed on LRE2017 dataset, it includes 5 major language clusters with 14 target dialects.
- Table below shows results on clean evaluation data in terms of accuracy in $\%$ (and Cavg in parenthesis).

Dur. (sec)	ivec [2]	LSTM [1]	HGRU
3	53.8 (0.53)	54.7 (0.55)	**55.1** (0.55)
10	72.3 (0.27)	72.1 (0.35)	**74.1** (0.32)
30	83.0 (0.13)	76.1 (0.28)	**83.0** (0.23)
1000	**56.2** (0.54)	42.8 (0.79)	53.5 (0.62)
overall	67.9 (0.37)	64.3 (0.48)	**68.5** (0.42)

Table 2: Results on clean LRE evaluation data

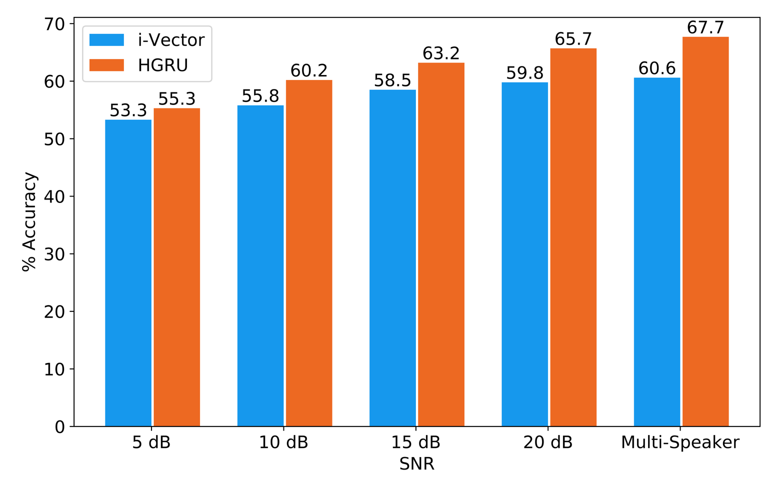


Figure 3: Partial noisy (10 sec.) and Multi speaker (3 sec. + 3 sec.) results

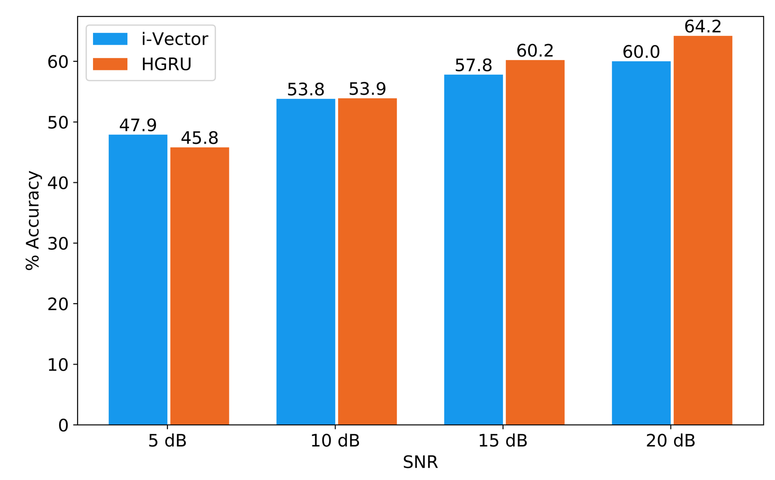
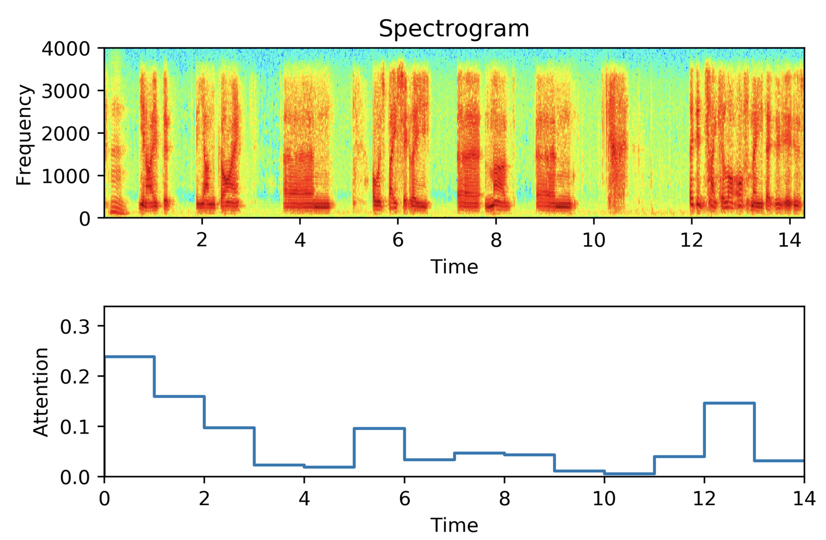


Figure 4: Noisy (10 sec.) results

- Comparable results when noise levels are high (5 dB and 10 dB SNR).
- **Significantly outperforms baseline when the audio has non-stationary characteristics like changing speaker or non-stationary noise levels.**

Attention Analysis

- In the transcription, green shade highlights the parts where attention was focused.
- **Vocalisations like 'aa', 'umm' were not given importance.**



Transcription:

0-3s : Umm you know……. young guy…

3s - 5s : ....ah....umm.... I.....

5s - 6s : I was basically...

6s - 12s : ah.... mugged…umm.................it did

12s - 13s : take a lot off me.....

13s - : stutter....

Figure 5: Attention on a clean British English audio file with transcript

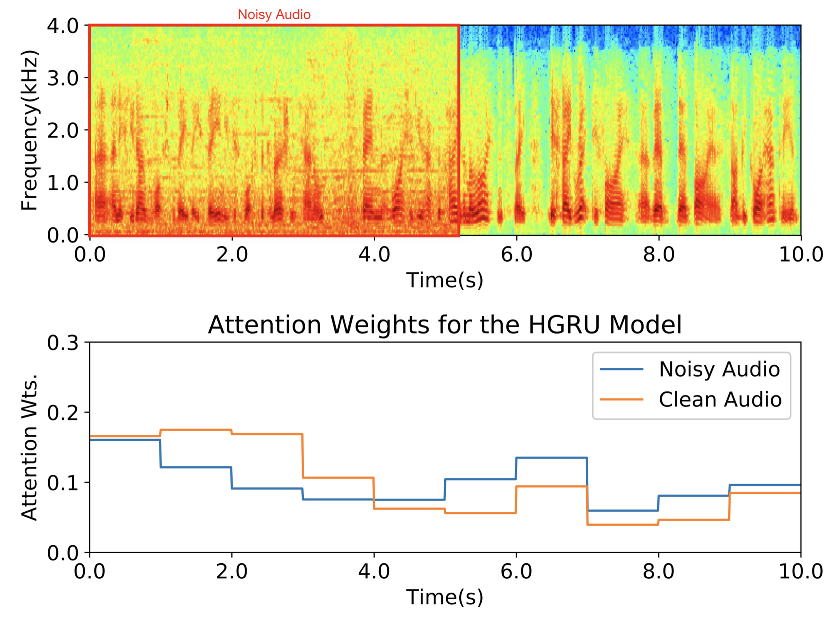


Figure 6: Attention weights of a partially noised audio file

- Noise (10 dB SNR) was added to the first 5 sec of the utterance to simulate non-stationary noisy environment.
- No preprocessing with speech activity detector.
- **HGRU was able to redistribute it's attention weights.**
- Attention weights **reduced in the noisy regions** while an increase in strength is observed in the cleaner regions.

Computational Complexity

	ivec [2]	LSTM [1]	HGRU
CPU	12	51	8
GPU	12	11.5	1.5

Table 3: Approximate computational time in seconds for ten 30sec eval files using a single CPU.

- **Architecture of HGRU allows for parallel computation unlike LSTM.**
- Noticeable improvement in the computational complexity achieved at comparable or improved LID performance.
- Machine Specification: 32 CPU, 8 core, 2 thread Intel x86-64 machine with 16 GB Nvidia Quadro P5000 GPU.

Summary

- Significantly improves over the previous attempts for end-to-end LSTM based language recognition systems [1].
- Robust to the presence of noise as well as in non-stationary conditions like partially corrupted speech data or multi-talker speech segments.
- The attention mechanism plays the role of relevance weighting.
- Low relative computational complexity.

Acknowledgements

- This work was funded partly by grants from the Department of Science and Technology (DST) Early Career Award (ECR01341) and the Pratiksha Young Investigator Award.
- We thank the Tata Trust, Indian Institute of Science, for the generous funding in support of ICASSP conference travel.

References

[1] Ruben Zazo, Alicia Lozano-Diez, and Joaquin Gonzalez-Rodriguez. Evaluation of an LSTM-RNN system in different nist language recognition frameworks. In *Proc. of Odyssey 2016 Speaker and Language Recognition Workshop*. ATVS-UAM, June 2016.

[2] Seyed Omid Sadjadi et al. The 2017 NIST language recognition evaluation. In *Proc. Odyssey*, Les Sables dÓlonne, France, June 2018.