Improving performance of Transformer based Low Resource Speech Recognition for Indian Languages

Vishwas M. Shetty, Metilda Sagaya Mary N J, S. Umesh

Indian Institute of Technology Madras, India
Address the Speech Recognition problem of Low Resource Indian Language using Transformers

Explore different ways of incorporating language information

- At character level.
- At acoustic feature level.

Our method of providing language information at feature level gave the best recognition performance.
Transformers in Speech Recognition
Transformers in ASR

Transformer

- An E2E framework based solely on attention.
- It comprises of a series of "Attention" (Self/Cross) and "Feed forward networks"
- Have shown promising results in several NLP tasks.

This has motivated its application to Automatic Speech Recognition.
Transformers in ASR
Dataset Details
## Dataset Details

<table>
<thead>
<tr>
<th>Language</th>
<th>Dur(hrs)</th>
<th>Train</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td></td>
<td>40</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td># Utter</td>
<td>22807</td>
<td>3075</td>
<td>3419</td>
</tr>
<tr>
<td>Tamil</td>
<td></td>
<td>40</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td># Utter</td>
<td>39131</td>
<td>3081</td>
<td>2609</td>
</tr>
<tr>
<td>Telugu</td>
<td></td>
<td>40</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td># Utter</td>
<td>44882</td>
<td>3040</td>
<td>2549</td>
</tr>
</tbody>
</table>
Transformers versus RNNs
Transformers versus RNNs

Models were trained based on Hybrid CTC/Attention approach using ESPNET tool on GTX 1080 GPU.

- RNN based model:
  - Four layer encoder with 320 BLSTM units
  - One layer decoder with 300 LSTM units
  - "Location" aware attention was used
  - Multi-task learning co-efficient - 0.5

- Transformer model
  - Twelve layer encoder with 2048 units
  - One layer decoder with 1024 units
  - Attention dimension of 256
  - Multi head attention with four attention heads
  - Multi-task learning co-efficient - 0.3
Transformers versus RNNs

<table>
<thead>
<tr>
<th>Language</th>
<th>Dev Set (Mono RNN)</th>
<th>Dev Set (Mono Trans)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>45.3</td>
<td>0</td>
</tr>
<tr>
<td>Tamil</td>
<td>39.6</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>43.9</td>
<td>0</td>
</tr>
</tbody>
</table>
Transformers versus RNNs

- Gujarati: Mono RNN 45.3, Mono Trans 31.9
- Tamil: Mono RNN 39.6, Mono Trans 34.1
- Telugu: Mono RNN 43.9, Mono Trans 36
Transformers versus RNNs

![Bar chart comparing Transformers and RNNs for Gujarati, Tamil, and Telugu languages on an Eval Set.](image)

- Gujarati: Mono RNN 50.2, Mono Trans 0
- Tamil: Mono RNN 38.8, Mono Trans 0
- Telugu: Mono RNN 45.6, Mono Trans 0
Transformers versus RNNs

The diagram compares the performance of Mono RNN and Mono Trans across three languages: Gujarati, Tamil, and Telugu. The evaluation set results are as follows:

- **Gujarati**: Mono RNN 50.2, Mono Trans 40.6
- **Tamil**: Mono RNN 38.8, Mono Trans 33.4
- **Telugu**: Mono RNN 45.6, Mono Trans 36.4
Can we do better than this?
Multilingual Models - Pool data from all the languages

- Build one single model by combining data from all the languages.
- The target character set is the union of characters from individual languages - 64 Guj, 48 Tam and, 64 Tel

<table>
<thead>
<tr>
<th>Language</th>
<th>Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>મેટો રેલનુ કમ્યુનિકેશન કેંદ્ર</td>
</tr>
<tr>
<td>Tamil</td>
<td>பிரதமப் புலவர்க்க</td>
</tr>
</tbody>
</table>
| Telugu    | సమ్మేధి తరపు ఇండియా మహామహామార్చి
Multilingual Model – Pooled data

- Gujarati
- Tamil
- Telugu

Mono Trans
Multi Trans

Dev Set

31.9
34.1
36
0 0 0
Multilingual Model – Pooled data

<table>
<thead>
<tr>
<th>Language</th>
<th>Mono Trans</th>
<th>Multi Trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>31.9</td>
<td>30.9</td>
</tr>
<tr>
<td>Tamil</td>
<td>34.1</td>
<td>34.7</td>
</tr>
<tr>
<td>Telugu</td>
<td>36</td>
<td>36.4</td>
</tr>
</tbody>
</table>
Multilingual Model – Pooled data

<table>
<thead>
<tr>
<th>Language</th>
<th>Eval Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>0</td>
</tr>
<tr>
<td>Tamil</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Mono Trans**: 40.6, 33.4, 36.4
- **Multi Trans**: 0, 0, 0
Multilingual Model – Pooled data

Gujarati: Mono Trans 40.6, Multi Trans 39.8
Tamil: Mono Trans 33.4, Multi Trans 34
Telugu: Mono Trans 36.4, Multi Trans 37
Can we make use of the language information?
Two possible Strategies

Assuming Language information known during training:

- Provide language information at the decoder – character level
- Provide language information at the encoder – feature level
Strategy 1: Provide Language information at the decoder

<table>
<thead>
<tr>
<th>Language</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td><code>&lt;guj_beg&gt; અતેકી તુમીલો તુમીલો &lt;guj_end&gt;</code></td>
</tr>
<tr>
<td>Tamil</td>
<td><code>&lt;tam_beg&gt; கெட்டாபாய் கெட்டாபாய் &lt;tam_end&gt;</code></td>
</tr>
<tr>
<td>Telugu</td>
<td><code>&lt;tlg_beg&gt; 欹્સ્લ્યો &lt;tlg_end&gt;</code></td>
</tr>
</tbody>
</table>

**Table 1**: Examples of LID in the target sequence

No LID used during testing – Hence **Universal Speech recognizer**.
Strategy 1: Provide Language information at the decoder

![Bar chart showing performance metrics for Gujarati, Tamil, and Telugu languages in the Dev Set.](chart)

- **Gujarati**:
  - Mono Trans: 31.9
  - Multi Trans: 0
  - LID Trans: 0

- **Tamil**:
  - Mono Trans: 34.1
  - Multi Trans: 0
  - LID Trans: 0

- **Telugu**:
  - Mono Trans: 36
  - Multi Trans: 0
  - LID Trans: 0
Strategy 1: Provide Language information at the decoder

![Bar chart showing performance metrics for Gujarati, Tamil, and Telugu languages using Mono Trans, Multi Trans, and LID Trans methods.](image)

- **Gujarati**
  - Mono Trans: 31.9
  - Multi Trans: 34.1
  - LID Trans: 0
- **Tamil**
  - Mono Trans: 30.9
  - Multi Trans: 34.7
  - LID Trans: 0
- **Telugu**
  - Mono Trans: 36
  - Multi Trans: 36.4
  - LID Trans: 0

**Legend**
- **Mono Trans**
- **Multi Trans**
- **LID Trans**

**Labels**
- **X-axis**: Dev Set
- **Y-axis**: Performance metric
Strategy 1: Provide Language information at the decoder

Gujarati
Tamil
Telugu
Mono Trans
Multi Trans
LID Trans

Dev Set
0
10
20
30
40

Gujarati
Tamil
Telugu

31.9 30.9 30
34.1 34.7 33.9
36 36.4 35.4
Strategy 1: Provide Language information at the decoder

![Bar chart showing evaluation set results for Gujarati, Tamil, and Telugu languages.]

- Gujarati: Mono Trans - 40.6, Multi Trans - 0, LID Trans - 0
- Tamil: Mono Trans - 33.4, Multi Trans - 0, LID Trans - 0
- Telugu: Mono Trans - 36.4, Multi Trans - 0, LID Trans - 0

Evaluations are shown on the x-axis as Gujarati, Tamil, and Telugu, and the y-axis represents the scores ranging from 0 to 50.
Strategy 1: Provide Language information at the decoder

![Bar chart showing evaluation set results for Gujarati, Tamil, and Telugu languages.](chart)

- **Gujarati**:
  - Mono Trans: 40.6
  - Multi Trans: 39.8
  - LID Trans: 0

- **Tamil**:
  - Mono Trans: 33.4
  - Multi Trans: 34
  - LID Trans: 0

- **Telugu**:
  - Mono Trans: 36.4
  - Multi Trans: 37
  - LID Trans: 0

Legend:
- Blue: Mono Trans
- Red: Multi Trans
- Yellow: LID Trans
Strategy 1: Provide Language information at the decoder

![Bar chart showing performance metrics for Gujarati, Tamil, and Telugu languages under different translation strategies. The chart compares mono-translations, multi-translations, and LID translations. The metrics are presented for each language and translation type.]
Strategy 2: Language information at the encoder

Provide language information at feature level:

- Append one hot vector
- Learn feature embeddings for the language

Language information is used while decoding, hence these models are **Language Specific Models**
Strategy 2a: Appending One hot vector
Strategy 2b: Learning Language Embedding
Strategy 2b: Learning Language embedding

- Analogous to how Character Embedding is learnt at the decoder
- Embedding vectors are initialized to a random vector with the dimension of acoustic feature
- Given an utterance - the targets belong to only one language - this information is used
Strategy 2: Results

![Bar chart showing results for Gujarati, Tamil, and Telugu languages in the Dev Set.]

- Gujarati: Mono Trans 31.9, Onehot Trans 0, Lang_embed Trans 0
- Tamil: Mono Trans 34.1, Onehot Trans 0, Lang_embed Trans 0
- Telugu: Mono Trans 36, Onehot Trans 0, Lang_embed Trans 0
Strategy 2: Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Mono Trans</th>
<th>Onehot Trans</th>
<th>Lang_embed Trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>31.9</td>
<td>29.4</td>
<td>0</td>
</tr>
<tr>
<td>Tamil</td>
<td>34.1</td>
<td>33.9</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>36.0</td>
<td>35.2</td>
<td>0</td>
</tr>
</tbody>
</table>
Strategy 2: Results

<table>
<thead>
<tr>
<th></th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev Set</td>
<td>31.9</td>
<td>34.1</td>
<td>36</td>
</tr>
<tr>
<td>Mono Trans</td>
<td>29.4</td>
<td>33.9</td>
<td>35.2</td>
</tr>
<tr>
<td>Onehot Trans</td>
<td>29.2</td>
<td>33.2</td>
<td>34.8</td>
</tr>
<tr>
<td>Lang_embed Trans</td>
<td>31.9</td>
<td>34.1</td>
<td>36</td>
</tr>
</tbody>
</table>

The bar chart compares the results of different translation methods for Gujarati, Tamil, and Telugu languages on the Dev Set.
Strategy 2: Results

![Bar Chart]

- **Mono Trans**
  - Gujarati: 40.6
  - Tamil: 33.4
  - Telugu: 36.4

- **Onehot Trans**
  - Gujarati: 0
  - Tamil: 0
  - Telugu: 0

- **Lang_embed Trans**
  - Gujarati: 0
  - Tamil: 0
  - Telugu: 0

Evaluator Set: 0, 10, 20, 30, 40, 50
Strategy 2: Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Eval Set</th>
<th>Mono Trans</th>
<th>Onehot Trans</th>
<th>Lang_embed Trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>0</td>
<td>40.6</td>
<td>38.3</td>
<td>0</td>
</tr>
<tr>
<td>Tamil</td>
<td>0</td>
<td>33.4</td>
<td>32.8</td>
<td>0</td>
</tr>
<tr>
<td>Telugu</td>
<td>0</td>
<td>36.4</td>
<td>35.9</td>
<td>0</td>
</tr>
</tbody>
</table>
Fine tuning
Fine tuning

Retrain well trained model using the target language

- Target Language
  - Targets
  - WELL TRAINED MODEL
  - Target Language

  - Multi Trans
  - LID Trans
  - Onehot Trans
  - Lang_embed Trans
Results after retraining

<table>
<thead>
<tr>
<th></th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>31.9</td>
<td>34.1</td>
<td>36</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Results after retraining

- Gujarati
  - Mono Trans: 31.9
  - Multi Retrain: 29.6
  - LID Retrain: 0
  - Onehot Retrain: 0
  - Lang_embed Retrain: 0

- Tamil
  - Mono Trans: 34.1
  - Multi Retrain: 33.1
  - LID Retrain: 0
  - Onehot Retrain: 0
  - Lang_embed Retrain: 0

- Telugu
  - Mono Trans: 36
  - Multi Retrain: 34.5
  - LID Retrain: 0
  - Onehot Retrain: 0
  - Lang_embed Retrain: 0

Dev Set: Gujarati, Tamil, Telugu
Results after retraining

Gujarati

- Mono Trans: 31.9
- Multi Retrain: 29.6
- LID Retrain: 29.3
- Onehot Retrain: 0
- Lang_embed Retrain: 0

Tamil

- Mono Trans: 34.1
- Multi Retrain: 33.1
- LID Retrain: 33
- Onehot Retrain: 0
- Lang_embed Retrain: 0

Telugu

- Mono Trans: 36
- Multi Retrain: 34.5
- LID Retrain: 34
- Onehot Retrain: 0
- Lang_embed Retrain: 0

Dev Set
Results after retraining

<table>
<thead>
<tr>
<th>Dev Set</th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>31.9</td>
<td>34.1</td>
<td>36</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>31.6</td>
<td>33.1</td>
<td>34.5</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>29.6</td>
<td>32.4</td>
<td>33.9</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td>29.3</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td>28.8</td>
<td>32</td>
<td>34</td>
</tr>
</tbody>
</table>

Languages: Gujarati, Tamil, Telugu
Results after retraining

<table>
<thead>
<tr>
<th></th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>31.9</td>
<td>34.1</td>
<td>36</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>29.6</td>
<td>33.1</td>
<td>34.5</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>29.3</td>
<td>32.4</td>
<td>33.9</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td>28.8</td>
<td>32.1</td>
<td>33</td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td>28.2</td>
<td>31.9</td>
<td>33</td>
</tr>
</tbody>
</table>

Dev Set

Gujarati

Tamil

Telugu
Results after retraining

- Gujarati: 40.6
- Tamil: 33.4
- Telugu: 36.4

Eval Set:
- Gujarati
- Tamil
- Telugu

Bars represent:
- Mono Trans
- Multi Retrain
- LID Retrain
- Onehot Retrain
- Lang_embed Retrain
Results after retraining

<table>
<thead>
<tr>
<th></th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>40.6</td>
<td>33.4</td>
<td>36.4</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>38.6</td>
<td>32.3</td>
<td>34.9</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Results after retraining

<table>
<thead>
<tr>
<th>Language</th>
<th>Eval Set</th>
<th>Mono Trans</th>
<th>Multi Retrain</th>
<th>LID Retrain</th>
<th>Onehot Retrain</th>
<th>Lang_embed Retrain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gujarati</td>
<td>40.6</td>
<td>38.6</td>
<td>38.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>33.4</td>
<td>32.3</td>
<td>32.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telugu</td>
<td>36.4</td>
<td>34.9</td>
<td>34.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Gujarati
- Tamil
- Telugu

- Eval Set: 0 0 0 0
Results after retraining

<table>
<thead>
<tr>
<th>Eval Set</th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>40.6</td>
<td>33.4</td>
<td>36.4</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>38.6</td>
<td>32.3</td>
<td>32.4</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>38.4</td>
<td>34.9</td>
<td>34.7</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results after retraining

<table>
<thead>
<tr>
<th>Eval Set</th>
<th>Gujarati</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Trans</td>
<td>40.6</td>
<td>38.6</td>
<td>36.4</td>
</tr>
<tr>
<td>Multi Retrain</td>
<td>38.4</td>
<td>33.4</td>
<td>31.6</td>
</tr>
<tr>
<td>LID Retrain</td>
<td>37.9</td>
<td>32.3</td>
<td>34.7</td>
</tr>
<tr>
<td>Onehot Retrain</td>
<td>36.9</td>
<td>32.4</td>
<td>34.7</td>
</tr>
<tr>
<td>Lang_embed Retrain</td>
<td>33.1</td>
<td>31.3</td>
<td>33.1</td>
</tr>
</tbody>
</table>
Conclusion
• Making use of Language information while training/decoding improves model performance.
• Explored two ways of incorporating Language information
• Providing Language information at the Encoder by learning feature embeddings, gave the best performance.
• Fine tuning further improved the model performance.