Sequence-based Multi-lingual Low Resource Speech Recognition

Siddharth Dalmia  Ramon Sanabria  Florian Metze  Alan W Black

sdalmia@cs.cmu.edu

Carnegie Mellon University
Language Technologies Institute
Language Universal
Multi-lingual Speech Recognition

• Many speech sounds are shared across languages.

• These sounds can be mapped to a set of language independent target units called International Phonetic Alphabet (IPA).

• However, these units are not always language agnostic.

Models trained with these units are prone to such errors!
Multiple target
Multi-lingual Speech Recognition

• What is an alternative?

• We can train a shared acoustic model with multiple targets, one for each language.

• The model learns to implicitly share the hidden space without the need of grounding them to language universal phonemes!
Multiple target
Multi-lingual Speech Recognition

• Lots of advantages
  
• Removes the need of having language universal phoneme set. They can even be characters of a language!

• We can use any of the existing datasets without preparing new labels or creating mappings of phonemes!
Previous Explorations

• Shared phone set with target language adaptation (T. Schultz et al, 2001; N. T. Vu et al, 2014)

• Language independent features like articulatory features (S. Stuker et al, 2003)

• Multilingual training of DNNs (A. Ghoshal et al, 2013; G. Heigold et al, 2013)

• Language-independent bottleneck features (K. Vesely et al, 2012; F. Grézl et al, 2014)

• Shared Phone Multilingual CTC Model (M. Müller, 2017)

• and many more…
CTC Based Multi-lingual ASR

• In this paper, we demonstrate that it is possible to train multi-lingual ASR directly on phone sequences and without explicitly using a shared phoneme set.

• We try to understand the effect of adding more languages (related or unrelated) in both multi-lingual and cross-lingual setting.

• We look into learning “bottleneck” like shared hidden acoustic representations and use it for cross-lingual adaptation.
Data - Babel Dataset

• We chose to perform experiments on a set of four languages which are the closest/have maximum phone overlap with Kurmanji.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Language</th>
<th>#Phones + Φ</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLing</td>
<td>Turkish</td>
<td>50</td>
<td>79 hrs</td>
</tr>
<tr>
<td></td>
<td>Haitian</td>
<td>40</td>
<td>67 hrs</td>
</tr>
<tr>
<td></td>
<td>Kazakh</td>
<td>70</td>
<td>39 hrs</td>
</tr>
<tr>
<td></td>
<td>Mongolian</td>
<td>61</td>
<td>46 hrs</td>
</tr>
</tbody>
</table>

• We test the effect of adding more languages by using SWBD (a large well prepared unrelated language) and BAB300 (a set of 4 unrelated languages in babel totaling to 300h).

• We do cross-lingual tests on Kurmanji (related) and Swahili (unrelated).
CTC Based Multi-lingual ASR

Model Parameters

Acoustic Model Params - 6 Layer BiLSTM with 360 hidden units.

WFST Params - Beam size of 9.0 and Lattice Beam of 4.0

Language Model - Lowest dev perplexity between 3-gram and 4-gram models.
Table 2: Word error rate (% WER) for each language in the MLing subset

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakh</th>
<th></th>
<th>Turkish</th>
<th></th>
<th>Haitian</th>
<th></th>
<th>Mongolian</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
</tr>
<tr>
<td>Monolingual</td>
<td>55.9</td>
<td>40.9</td>
<td>53.1</td>
<td>36.2</td>
<td>49.0</td>
<td>36.9</td>
<td>58.2</td>
<td>45.2</td>
</tr>
</tbody>
</table>

CTC Based Multi-lingual ASR

BASELINE
CTC Based Multi-lingual ASR
It works!

Table 2: Word error rate (% WER) for each language in the MLing subset

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakh</th>
<th>Turkish</th>
<th>Haitian</th>
<th>Mongolian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
</tr>
<tr>
<td>Monolingual</td>
<td>55.9</td>
<td>40.9</td>
<td>53.1</td>
<td>36.2</td>
</tr>
<tr>
<td>Multilingual</td>
<td>53.2</td>
<td>36.5</td>
<td>52.8</td>
<td>34.4</td>
</tr>
</tbody>
</table>

~1.5 % WER↓

Multi-lingual Training
CTC Based Multi-lingual ASR Improves further!

Table 2: Word error rate (% WER) for each language in the MLing subset

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakh</th>
<th>Turkish</th>
<th>Haitian</th>
<th>Mongolian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
</tr>
<tr>
<td>Monolingual</td>
<td>55.9</td>
<td>40.9</td>
<td>53.1</td>
<td>36.2</td>
</tr>
<tr>
<td>Multilingual</td>
<td>53.2</td>
<td>36.5</td>
<td>52.8</td>
<td>34.4</td>
</tr>
<tr>
<td>+ FineTuning</td>
<td>50.6</td>
<td>35.1</td>
<td>49.0</td>
<td>32.2</td>
</tr>
</tbody>
</table>

Fine-tuning for each language

~4 % WER↓

Note: Improvements are higher for lower resourced languages!
Table 2: Word error rate (% WER) for each language in the MLing subset

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakh</th>
<th>Turkish</th>
<th>Haitian</th>
<th>Mongolian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>PER</td>
<td>WER</td>
<td>PER</td>
</tr>
<tr>
<td>Monolingual</td>
<td>55.9</td>
<td>40.9</td>
<td>53.1</td>
<td>36.2</td>
</tr>
<tr>
<td>Multilingual</td>
<td>53.2</td>
<td>36.5</td>
<td>52.8</td>
<td>34.4</td>
</tr>
<tr>
<td>+ FineTuning</td>
<td>50.6</td>
<td>35.1</td>
<td>49.0</td>
<td>32.2</td>
</tr>
<tr>
<td>Multilingual + SWBD</td>
<td>52.3</td>
<td>36.6</td>
<td>51.3</td>
<td>33.0</td>
</tr>
<tr>
<td>+ FineTuning</td>
<td>48.2</td>
<td>33.5</td>
<td>48.7</td>
<td>31.9</td>
</tr>
</tbody>
</table>

What if you add English Switchboard (300h) !?

Multi-lingual Training with SWBD and Fine-Tuning

~6 % WER↓
### SWBD vs Bab300

- Using 300 hours of various Babel languages performs worse than just adding SWBD.

**Table 2: Word error rate (% WER) on the test languages.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakh</th>
<th>Turkish</th>
<th>Haitian</th>
<th>Mongolian</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLing + Bab300</td>
<td>57.5</td>
<td>52.0</td>
<td>47.8</td>
<td>56.7</td>
</tr>
<tr>
<td>MLing + SWBD</td>
<td>52.3</td>
<td>51.3</td>
<td>45.8</td>
<td>54.5</td>
</tr>
</tbody>
</table>

- It is beneficial to add large amounts of well-prepared data from a single language rather than adding many unrelated languages.

- Adding a large number of languages may in fact prevent the model from training well.
Can this layer be used as a **discriminatory audio feature layer** that is independent of the input language?

Motivation from bottleneck layers!
We take the encoder representations of various trained model.

Then train only the softmax layer using various amounts of data from a related unseen language, Kurmanji.

Check for whether the pre-trained hidden representation can linearly separate a new language into it’s phoneme sequence.
Representation Learning

Monolingual Turkish model performs much better than SWBD.
Multilingual Models are considerably better than the Monolingual models!
Using more languages help!
Better language independent hidden representation

BAB300 does better than SWBD!
Representation Learning

Multilingual models by JUST using 10-20% data to ONLY adapt the softmax layer => almost close to a Monolingual Kurmanji Model on 100% data with ALL layers trained.
Multilingual system surpasses the mono-lingual baseline when just 25% of the original data has been seen!
This behavior of retraining ("full network adaptation") seems independent of the target language.
Kurmanji performs well, because the language is similar to the training languages.

Larger gap while adapting to an unrelated language, in this case Swahili.
Cross-lingual Explorations

Initialization with many languages (MLing + BAB300) is beneficial!
Cross-lingual Explorations

When the entire network can be retrained -
Starting with (MLing + SWBD) or (MLing + BAB300) perform almost equally well!
Cross-lingual Explorations

Full Network adaptation (on the right) outperforms Softmax adaptation (on the left) as soon as 2-4 h of data become available.
In very low resource cross-lingual scenarios, it is probably better to adapt a model to an unseen language by re-training the softmax layer.
Conclusion

• It is possible to train multi-lingual and cross-lingual acoustic models directly on phone sequences.

• These models can learn a language independent representation.

• In multi-lingual settings, it seems beneficial to train on related languages only, or on large amounts of clean data.
Conclusion

• In very low resource cross-lingual scenarios, training on related languages help, as does training on many languages, rather than large amounts of single language.

• The effect of the choice of languages disappears as more and more data is available and the whole network can be retrained.
Steps Ahead

• Can we do ASR on a language without any training data?

• Use a language universal recognizer (shared softmax layer).

• Decode using a phoneme based neural language models trained on nonparallel text.

• Thereby facilitating us to do zero-resource speech recognition!
Thank you!

• Code available in -
  https://github.com/srvk/eesen/tree/tf_clean/asr_egs/babel/105_201_302_401

• Contact us - {sdalmia,ramons,fmetze,awb}@cs.cmu.edu

Questions?