Lattice-Free Maximum Mutual Information Training of Multilingual Speech Recognition Systems

Srikanth Madikeri\textsuperscript{1}, Banriskhem K. Khonglah\textsuperscript{1}, Sibo Tong \textsuperscript{1,2}, Petr Motlicek\textsuperscript{1}, Hervé Bourlard\textsuperscript{1,2}, Daniel Povey\textsuperscript{3}

\textsuperscript{1}Idiap Research Institute, Martigny, Switzerland
\textsuperscript{2}Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
\textsuperscript{3}Xiaomi Corporation, Beijing, China

msrikanth, bkhonglah, stong, petr.motlicek, bourlard@idiap.ch, dpovey@xiaomi.com

Abstract

Multilingual acoustic model training combines data from multiple languages to train an automatic speech recognition system. Such a system is beneficial when training data for a target language is limited. Lattice-Free Maximum Mutual Information (LF-MMI) training performs sequence discrimination by introducing competing hypotheses through a denominator graph computing the probability over all possible sequences based on ground truth and a denominator graph. The standard approach to train a multilingual model with LF-MMI is to combine the acoustic units from all languages and use a common denominator graph. The resulting model is either used as a feature extractor to train an acoustic model for the target language or directly fine-tuned. In this work, we propose a scalable approach to train the multilingual acoustic model using a typical multitask network for the LF-MMI framework. A set of language-dependent denominator graphs is used to compute the cost function. The proposed approach is evaluated under typical multilingual ASR tasks using GlobalPhone and BABEL datasets. Relative improvements up to 13.2\% in WER are obtained when compared to the corresponding monolingual LF-MMI baselines. The implementation is made available as a part of the Kaldi speech recognition toolkit.

Index Terms: speech recognition, multilingual ASR, LF-MMI

1. Introduction

In Automatic Speech Recognition (ASR) for low-resource languages, training multilingual systems is an effective way to compensate for limited amount of data \cite{1, 2, 3, 4, 5, 6}. When trained with resources from multiple languages, Deep Neural Networks (DNN) based Acoustic Models (AM) can function as a feature extractor to train a monolingual acoustic model for the target language \cite{7, 8, 9}. Alternately, the models can be adapted to the target language \cite{10, 11, 12, 13, 14, 15}. The multilingual models can either share the output layer or have separate output layers (one for each language) \cite{3}. In the former case, monophones may be used to avoid a huge output layer, which is often followed by retraining the network for the target language with senones.

In this work, we focus on sequence-discriminative training of multilingual AM with the Lattice-Free Maximum Mutual Information (LF-MMI) framework \cite{16}. LF-MMI training has been shown to have superior performance compared to the conventional cross-entropy (CE) training of DNNs \cite{17, 18}. The MMI cost function uses a numerator graph modelling the observed feature sequence based on ground truth and a denominator graph computing the probability over all possible sequences \cite{19}. The latter enforces the discriminative property in the training shown to be useful for training AM \cite{20, 21}.

Given the advantages of both multilingual and LF-MMI training procedures, it is natural to combine them to obtain better performance. In \cite{22, 23}, multilingual LF-MMI models were observed to improve over their monolingual counterparts. The multilingual resources are combined by merging the phoneme sets from all languages either using a universal phone set such as the International Phonetic Alphabet (IPA), or by combining acoustic units. In both cases, a universal denominator graph is shared across all languages during training.

When combining acoustic units for multilingual training, the output layer size increases rapidly with number of languages. This may render such a system impractical during decoding. We refer to this type of multilingual AM as a single-task system. Alternately, multitask training solves this issue by separating the output layers of languages so that during decoding only the output relevant to the language is used. An added advantage during training is that the cost function can be computed faster as its complexity depends on the number of states in the denominator.

In this paper, we compare different styles of multilingual training in the LF-MMI framework. The two styles are broadly categorized as single-task and multitask depending on whether the output layer is shared across languages or not. For single-task training, existing LF-MMI implementation can be easily extended. For multitask training, we make our implementation available as a part of Kaldi \cite{24}. The comparisons are performed on two commonly used multilingual databases: (1) GlobalPhone and (2) BABEL. We present results on 5 target languages for the former and 4 target languages for the latter. The results show that multitask training provides a much more scalable approach to develop multilingual AM due to the aforementioned advantages without any loss in performance. The rest of the paper is organized as follows: in Section 2, the LF-MMI training procedure is described. In Section 3, the proposed multilingual LF-MMI training procedure is given. Results of our experiments on GlobalPhone and BABEL are described in Section 4.

2. LF-MMI

In LF-MMI, the MMI objective function is used as the cost function to train the AM \cite{16}. The cost function is given as

\footnotesize
\begin{align*}
\end{align*}
\normalsize

\begin{align}
\text{http://dx.doi.org/10.21437/Interspeech.2020-2919}
\end{align}
The numerator graph is constructed using forced alignment and the denominator graph is obtained by composing the phone language model with the phonetic context-dependency followed during training is limited.

While training a multilingual model with the output layer containing acoustic units from all languages, the objective still remains the same as above, meaning the $\text{DEN}_{\gamma}$ is language-independent.

Computing $\text{DEN}_{\gamma}$ requires training a phone language model. Combining acoustic units across all languages for single-task multilingual training not only increases the number of states in the denominator graph, but may also introduce lead to noisy $\text{DEN}_{\gamma}$ estimates. Thus, to reduce the influence of other languages while computing $\text{DEN}_{\gamma}$, we propose to use a set of language-dependent denominator for AMs trained in multitask fashion.

### 3. Multilingual LF-MMI

Multiple approaches exist to train a multilingual AM. Depending on whether the output layer is shared by languages or not, we can classify it as either single-task or multitask model. The difference between these two broad categories of multilingual LF-MMI systems is shown in Figure 1. In the multitask architecture, each language has a separate output layer preceded by a pre-final layer and a corresponding objective function (marked $F_{\text{MMI}}^{(1)} \ldots F_{\text{MMI}}^{(L)}$ in the figure).

The choice between single-task and multitask AM dictates how the acoustic units are shared across languages. In the single-task case, one can simply combine the acoustic units by choosing a union of all non-silence acoustic units from each language. Alternately, well-defined linguistic units such as IPA can be used to derive the acoustic units. In the multitask case, each language will have its own set of acoustic units.

Given such possibilities to train the AM, the single-task configuration also provides a choice of using language-specific (i.e. trained with data from all languages) or language-independent denominator (i.e. trained with data from only one language), whereas only the language-independent denominator is applicable in the multitask case. The focus of this paper is to compare all such possible configurations to better understand the performance of the resulting models.

In single-task multilingual AM, the case of using language-independent denominator is equivalent to training monolingual AMs. However, when using language-specific denominators, the cost function changes as follows: we have $L$ objective func-
tions, where $L$ is the number of languages, computed independent of each other depending only on the language of the utterance:

$$\mathcal{F}_{\text{MMI}}^{(l)} = \frac{1}{U} \sum_{u=1}^{U} \log \frac{p(\mathbf{x}(u) | \mathcal{M}_{\text{w}(u)}^{l}, \theta) p(\mathbf{w}(u))}{p(\mathbf{x}(u) | \mathcal{M}_{\text{den}}^{l}, \theta)},$$  \tag{3}

where $U$ is the number of utterances in the minibatch for language $l$, $\theta$ contains the shared and language-dependent parameters, $\mathcal{M}_{\text{w}(u)}^{l}$ and $\mathcal{M}_{\text{den}}^{l}$ are language-specific numerator and denominator graphs, respectively.

Each denominator graph is built from the language-specific phone language model (the same as that used in monolingual LF-MMI training). Gradients for language-dependent layers are computed and updated for each minibatch. Using backpropagation, the shared parameters are then updated. The overall cost function is the weighted sum of all language-dependent cost functions:

$$\mathcal{F}_{\text{MMI}} = \sum_{l=1}^{L} \alpha_l \mathcal{F}_{\text{MMI}}^{(l)},$$  \tag{4}

where $\alpha_l$ is language-dependent weight. Note that each minibatch is expected to have samples (sequence of MFCCs) from multiple languages. To facilitate such a training in Kaldi, we modify the training procedure to select the denominator graph for each sequence in the minibatch according to the language. In practice, this only requires the knowledge of the language of each sequence in the minibatch. Assuming the sequences are grouped by language, we simply iterate over the languages in the minibatch to call the existing procedures for monolingual training with the appropriate denominator graph. Such multitask models also simplify addition or removal of languages and applying language-specific operations during training.

4. Experiments

Experiment results are reported on GlobalPhone [25] and BABEL datasets. All experiments are performed with the Kaldi toolkit [24]. For GlobalPhone, we used the French (FR), German (GE), Portuguese (PO), Russian (RU) and Spanish (SP) datasets from the GlobalPhone corpus [23]. Each language has roughly 20 hours of speech for training and two hours for development and evaluation sets, from a total of about 100 speakers. The development sets were used to tune the hyper-parameters for training. Only the results on evaluation sets are reported. The trigram language models that we used are publicly available\(^2\). The detailed statistics for each of the languages is given in Table 1.

We also investigated our proposed method with the BABEL dataset. Datasets for several languages with limited resources were released during the BABEL project with the main goal of building keyword spotting systems. We considered 4 BABEL languages for evaluation: Tagalog (TGL), Swahili (SWA), Zulu (ZUL), and Turkish (TUR). The statistics of the target languages are given in Table 2. Trigram language models are used during testing.

4.1. GlobalPhone Setup

We used 40-dimensional MFCCs as acoustic features, derived from 25 ms frames with a 10 ms frame shift. The features were normalized via mean subtraction and variance normalization on a speaker basis. We used a frame subsampling factor of 3 which speeds up training by a factor of 2. We also augmented the data with 2-fold speed perturbation in all the experiments. The network consists of 8 layers of Time Delay Neural Network (TDNN), with 450 nodes in each layer [26].

The network consists of 8 layers of Time Delay Neural Network (TDNN), with 450 nodes in each layer [26].

4.2. GlobalPhone Results

The results on GlobalPhone are presented in Table 3. The single-task multilingual systems trained with concatenated phone set improve over the monolingual LF-MMI systems on four out of five languages. Using language-dependent denominator, in this case, does not make a significant difference in terms of WERs, thus only providing computational benefits during training. The single-task system performs better on FR and GE than the multitask system. The difference on the other languages is marginal. The multitask multilingual system improves over the monolingual baseline for 4 out of 5 languages. The relative improvements range from 0.7% (for PO) to 10% (for RU). We do not compare to the CE system as its results are poorer compared to the two LF-MMI baselines. We believe that the LF-MMI baselines are superior due to the controlled nature of the dataset (read speech and clean acoustic conditions).

4.3. BABEL setup

We consider two training configurations: training with only 4 of the target languages and training with 14 languages. The 14-language system is used to demonstrate the scalability of the

\(^2\)http://www.csl.uni-bremen.de/GlobalPhone/

Table 1: Statistics of the subset of GlobalPhone languages used in this work: the amounts of speech data for training and evaluation sets are in hours.

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocabulary</th>
<th>PPL</th>
<th>#Phones</th>
<th>Train (h)</th>
<th>Dev (h)</th>
<th>Eval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>65k</td>
<td>324</td>
<td>38</td>
<td>22.7</td>
<td>2.1</td>
<td>2.0</td>
</tr>
<tr>
<td>GE</td>
<td>38k</td>
<td>672</td>
<td>41</td>
<td>14.9</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>PO</td>
<td>62k</td>
<td>58</td>
<td>45</td>
<td>22.7</td>
<td>1.6</td>
<td>1.8</td>
</tr>
<tr>
<td>RU</td>
<td>293k</td>
<td>1310</td>
<td>48</td>
<td>21.1</td>
<td>2.7</td>
<td>2.4</td>
</tr>
<tr>
<td>SP</td>
<td>19k</td>
<td>154</td>
<td>40</td>
<td>17.6</td>
<td>2.0</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 2: Statistics of BABEL target languages used for testing. Note that the Eval sets mentioned refer to the “dev” set in the official BABEL release. Only conversational speech is considered for both training and testing. All durations are calculated prior to silence removal. (PPL: perplexity)

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocabulary</th>
<th>PPL</th>
<th>#Phones</th>
<th>Train (h)</th>
<th>Eval (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagalog</td>
<td>22k</td>
<td>148</td>
<td>48</td>
<td>84.5</td>
<td>10.7</td>
</tr>
<tr>
<td>Swahili</td>
<td>25k</td>
<td>357</td>
<td>38.0</td>
<td>37.0</td>
<td>9.3</td>
</tr>
<tr>
<td>Turkish</td>
<td>41k</td>
<td>396</td>
<td>77.2</td>
<td>96.0</td>
<td>9.4</td>
</tr>
<tr>
<td>Zulu</td>
<td>56k</td>
<td>719</td>
<td>56.7</td>
<td>57.0</td>
<td>9.2</td>
</tr>
</tbody>
</table>
multitask system. In both cases, results for only 4 target languages are reported (see Table 4). We follow the feature configuration (except for feature mean and variance normalization) and data augmentation of GlobalPhone systems. In addition, an online i-vector extractor of dimension 100 is trained for each configuration. The transcripts are used for speech/non-speech labels. The online i-vectors are appended to MFCCs as input to the DNN. TDNN architecture is used with 8 hidden layers. Each hidden layer has 1024 units. The pre-final layer has only 200 units. Frame-dropping is enabled for all models.

In order to obtain alignments to train all the TDNN models, HMM/GMM models were first trained for each language. The standard recipe from Kaldi was followed.

**5. Conclusions**

In this work, we compared different styles of training multilingual acoustic models in the LF-MMI framework. The system was evaluated on GlobalPhone and BABEL datasets. The results on target languages in GlobalPhone show that the multitask training approach leads to a system that outperforms single-task models trained with either IPA or combined phone sets. The results on BABEL datasets show similar trends in improvement for 3 out of 4 target languages. By further increasing the number of languages in training significant benefits are achieved demonstrating the scalability of our method. We obtained relative improvements up to 13.2% when compared to the monolingual model.

**6. Acknowledgement**

The research is based upon the work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The work is also partially supported by the ROXANNE H2020 EC project (http://www.roxanne-euproject.org), under grant agreement No. 833635.

---

### Table 3: Comparison between target languages in GlobalPhone in WER(%) (FR: French, GE: German, PO: Portuguese, RU: Russian, SP: Spanish)

<table>
<thead>
<tr>
<th>System</th>
<th>FR</th>
<th>GE</th>
<th>PO</th>
<th>RU</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual LF-MMI</td>
<td>20.4</td>
<td>12.7</td>
<td>15.2</td>
<td>24.6</td>
<td>7.1</td>
</tr>
<tr>
<td>Single-task multilingual system</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language independent</td>
<td>21.3</td>
<td>12.5</td>
<td>14.9</td>
<td>22.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Language dependent</td>
<td>21.3</td>
<td>12.4</td>
<td>15.0</td>
<td>22.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Multitask multilingual system</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 languages</td>
<td>20.7</td>
<td>11.7</td>
<td>15.1</td>
<td>22.1</td>
<td>6.5</td>
</tr>
</tbody>
</table>

### Table 4: BABEL languages used for training and testing.

<table>
<thead>
<tr>
<th>Category</th>
<th>Languages (except for feature mean and variance normalization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target languages &amp; 4 Language Training</td>
<td>Tagalog, Swahili, Zulu, Turkish &amp; 4 Language Training</td>
</tr>
<tr>
<td>14 Language Training</td>
<td>Tagalog, Swahili, Zulu, Turkish, Assamese, Bengali, Cantonese, Haitian, Kazhak, Kurmanji, Tamil, Telugu, Tok, Vietnamese</td>
</tr>
</tbody>
</table>

### Table 5: Comparison between target languages in BABEL in WER(%) Improvements with LF-MMI are in bold. (TGL: Tagalog, SWA: Swahili, TUR: Turkish, ZUL: Zulu)

<table>
<thead>
<tr>
<th>System</th>
<th>TGL</th>
<th>SWA</th>
<th>TUR</th>
<th>ZUL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual LF-MMI</td>
<td>45.3</td>
<td>38.7</td>
<td>47.2</td>
<td>53.5</td>
</tr>
<tr>
<td>Single-task multilingual system</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language independent</td>
<td>44.4</td>
<td>35.5</td>
<td>43.4</td>
<td>52.4</td>
</tr>
<tr>
<td>Language dependent</td>
<td>44.4</td>
<td><strong>35.4</strong></td>
<td><strong>43.0</strong></td>
<td>51.9</td>
</tr>
</tbody>
</table>

### 5. Acknowledgement

The research is based upon the work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The work is also partially supported by the ROXANNE H2020 EC project (http://www.roxanne-euproject.org), under grant agreement No. 833635.
7. References


