AN INVESTIGATION OF MULTILINGUAL ASR USING END-TO-END LF-MMI

Sibo Tong[1,2]*, Philip N. Garner*[1]*, Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

The end-to-end lattice-free maximum mutual information (LF-MMI) approach has recently been shown to be beneficial for automatic speech recognition (ASR) in general. More specifically, its end-to-end nature and use of context independent phone labels make it attractive for multilingual ASR. We show that end-to-end LF-MMI is indeed competitive on a low-resourced multilingual task, comfortably outperforming a connectionist temporal classification (CTC) baseline. We further investigate the feasibility of biphone contexts, being a candidate compromise between the context independent approach and the triphone contexts that usually perform well. We show that biphones do not initially perform well, but can do so after language adaptive training, concluding that biphones carry language variability but are promising for multilingual ASR.

Index Terms— end-to-end LF-MMI, multilingual ASR, CTC, language adaptive training

1. INTRODUCTION

Recently, there has been increased interest in rapidly developing high performance automatic speech recognition (ASR) systems for a broad range of languages. Speech recognition systems built with multilingual deep neural networks (DNNs) have been shown to provide consistent advantages especially for low-resourced languages [1, 2, 3]. In DNN, the hidden layers can be considered as a universal feature extractor. Therefore, the hidden layers can be trained jointly using data from multiple languages to benefit each other [3, 4]. The target of the multilingual DNN can be either the universal International Phonetic Alphabet (IPA) based multilingual senones [5] or a layer of separate activations for each language [3, 6, 7].

All of these models are based on a conventional DNN-HMM framework. In order to perform well, DNNs model context-dependent states to mitigate the error associated with the Markov assumption. However, this creates more challenges for multilingual and cross-lingual ASR because of the large increase in context dependent labels arising from the phone set mismatch. Recently, end-to-end approaches for automatic speech recognition have received a lot of attention. Popular end-to-end approaches are Connectionist Temporal Classification (CTC) [8], RNN-Transducers [9] and attention-based methods [10]. More recently, end-to-end lattice-free maximum mutual information (LF-MMI) has been proposed [11]. These methods typically aim to train a neural network-based acoustic model in one stage without relying on prerequisite models, alignments or decision trees. Multilingual ASR and cross-lingual adaptation can benefit more from these properties: language-specific prerequisite systems are no longer required; cross-lingual adaptation from an IPA-based system can be done simply by extending the output layer to new phonemes in a target language [12]. CTC training has been shown to be a promising alternative to the traditional DNN-HMM system for both multilingual ASR and cross-lingual adaptation [12, 13].

CTC-based models, however, are sensitive to the amount of training data. Recently, it was shown that the end-to-end LF-MMI training can achieve comparable or better performance than other end-to-end approaches, including CTC, on well-known large vocabulary tasks. Therefore, it is natural to investigate the performance of the end-to-end LF-MMI in low-resourced scenarios. Although several prior works have applied regular LF-MMI (not the end-to-end version) in multilingual training [14, 15], most of them used the multilingual network as a seed model for further transfer learning. To the best of our knowledge, this is the first work investigating end-to-end LF-MMI for multilingual ASR.

To this end, we first discuss and compare CTC and end-to-end LF-MMI in Section 2, as CTC was a pioneering approach in end-to-end speech recognition and was shown to be a promising alternative for cross-lingual adaptation. Then, the IPA-based universal multilingual training approach is described in Section 3. We investigate biphone modelling, finding that it is sensitive to variability across languages, and hence requires some language adaptation in order to work. Therefore, language adaptive training (LAT) is introduced and investigated in the context of end-to-end LF-MMI training. Moreover, a pruned biphone tree is also proposed to remove the redundant cross-lingual combinations in Section 3. Experimental results and analysis are provided in Section 4. Finally, Section 5 concludes the paper.

2. END-TO-END MODELS

2.1. Connectionist Temporal Classification

The Connectionist Temporal Classification (CTC) approach uses an objective function for sequence labelling problems without requiring any frame-level alignment between the input and target labels. For an input sequence $\mathbf{X}^{(u)} = (\mathbf{x}_1^{(u)}, \ldots, \mathbf{x}_{T_u}^{(u)})$, the conditional probability $P(\mathbf{w}^{(u)}|\mathbf{X}^{(u)}, \theta)$ is obtained by summing over the probabilities of all the paths that correspond to the target label sequence $\mathbf{w}^{(u)}$ after inserting the repetitions of labels and the blank tokens, i.e.,

$$\mathcal{F}_{CTC} = \sum_{u=1}^{U} \log p(\mathbf{w}^{(u)}|\mathbf{x}^{(u)}, \theta) \quad (1)$$

$$= \sum_{u=1}^{U} \log \sum_{\mathbf{s} \in \Omega(\mathbf{w}^{(u)})} \prod_{t=1}^{T_u} p(s_t|\mathbf{x}_t^{(u)}, \theta) \quad (2)$$

where $\Omega(\mathbf{w}^{(u)})$ denotes the set of all possible paths that correspond to $\mathbf{w}^{(u)}$ after repetitions of labels and insertions of the blank token and θ represents the model parameters. The conditional probability of the labels at each time step, $P(s_t|\mathbf{x}_t, \theta)$, is estimated using a neural network. More details can be found in [8].

As formulated by [16], CTC can be identified as a special case of the generalized hybrid HMM/NN training procedure using the full-sum over the hidden state sequence. The generalized HMM training optimizes the likelihood of observing $\mathbf{x}^{(u)}$ given a target sequence $\mathbf{w}^{(u)}$ with state sequences $\mathbf{s}$ as hidden variable and model parameters θ, given by:

$$\mathcal{F}_{ML} = \sum_{u=1}^{U} \log p(\mathbf{x}^{(u)}|\mathbb{M}_{\mathbf{w}^{(u)}}, \theta) \quad (3)$$

$$= \sum_{u=1}^{U} \log \sum_{\mathbf{s} \in \mathbb{M}_{\mathbf{w}^{(u)}}} \prod_{t=1}^{T_u} p(s_{t+1}|s_t) p(\mathbf{x}_t^{(u)}|s_t, \theta) \quad (4)$$

where the composite HMM graph $\mathbb{M}_{\mathbf{w}^{(u)}}$ represents all the possible state sequences $\mathbf{s}$ pertaining to the transcription $\mathbf{w}^{(u)}$. In HMM/NN models, $p(\mathbf{x}_t|s_t, \theta)$ is modeled as

$$p(\mathbf{x}_t|s_t, \theta) \sim \frac{p(s_t|\mathbf{x}_t, \theta)}{p(s_t)} \quad (5)$$

In this context, CTC can be considered as a special reduced HMM topology which has no transition probabilities, no state prior probability model but a special blank state and is trained with Baum-Welch soft alignments.

2.2. End-to-end LF-MMI

Maximum mutual information (MMI) is a discriminative objective function which aims to maximize the probability of the reference transcription, while minimizing the probability of all other transcriptions:

$$\mathcal{F}_{MMI} = \sum_{u=1}^{U} \log \frac{p(\mathbf{x}^{(u)}|\mathbb{M}_{\mathbf{w}^{(u)}}, \theta)}{p(\mathbf{x}^{(u)})} \quad (6)$$

$$= \sum_{u=1}^{U} \log \frac{p(\mathbf{x}^{(u)}|\mathbb{M}_{\mathbf{w}^{(u)}}, \theta)}{\sum_{\mathbf{w}} p(\mathbf{x}^{(u)}|\mathbb{M}_{\mathbf{w}}, \theta)} \quad (7)$$

In the regular LF-MMI proposed in [17], the composite HMM was not used as the numerator graph and instead a special acyclic graph was used which could exploit the alignment information from a previous HMM-GMM model. By contrast, in the end-to-end LF-MMI proposed in [11], the composite HMM (with self-loops) was used as the numerator graph. As a result, unlike regular LF-MMI, there is no prior alignment information in the numerator graph and there is no restriction on the self-loops so there is much more freedom for the neural network to learn the alignments. Comparing (1), (3) and (6), we can consider the end-to-end LF-MMI as a discriminative version of CTC training.

3. MULTILINGUAL PHONEME-BASED MODEL

3.1. Universal Phone Set

More recently, building end-to-end multilingual speech recognition systems using a universal grapheme set has been investigated [18, 19]. However, modelling graphemes includes implicit modelling of spelling, which requires a large amount of data. Moreover, graphemes can differ a lot from language to language. Languages that have nothing in common in terms of graphemes also share some common phonemes. Moreover, a universal phoneme-based model is easily extensible to unseen phonemes when adapted to a new language [12].

With this motivation, and following our previous work[12, 13], we propose a multilingual architecture that uses a universal output label set consisting of the union of all phonemes from the multiple languages. This universal phone set can be either derived in a data-driven way, or obtained from the International Phonetic Alphabet (IPA). In this study, we created a universal phone set by merging the monolingual phones which share the same symbol in the IPA table.

For multilingual end-to-end LF-MMI training, we trained a multilingual phoneme language model for denominator graph using the training transcriptions from all the multilingual data. The composite HMM graphs were created using the language-specific lexicons, and were used as the numerator graphs. In this sense, the numerator graph is language-specific while the denominator graph is multilingual.

3.2. Biphone Modelling and Pruned Biphone Tree

Although monophone-based end-to-end training fits well for multilingual ASR because of its simplicity, it is well known

that context-dependent modelling further improves the performance. In this sense, using full biphones can be a good compromise. It has been shown that context-dependent modelling also helps in end-to-end LF-MMI training [11]. This was implemented as a trivial full biphone tree. This tree is not pruned at all and does not have any tying, so there is no need for alignments and the approach does not require any previously trained models. However the size of the biphone targets grows quadratically in a multilingual set-up. A lot of cross-lingual biphone combinations will be created which never occur in the training data, impacting the training efficiency. Therefore, we propose to build a pruned biphone tree where all the cross-lingual biphone combinations are pruned away. More specifically, suppose a language has a phone set of $\{a, b, c\}$ and another language has a phone set of $\{b, c, d\}$. The universal phone set would be $\{a, b, c, d\}$. When creating the biphone targets, combinations such $a - d$ and $d - a$ will also be generated. However, they will never appear in the training data and are pruned away in this work.

3.3. Language Adaptive Training

It has been demonstrated that the layers close to the output layer are more language-related and training the last layer in a language-dependent manner can help the IPA-based multilingual system to better capture the language specificity [20]. More specifically, the output of the neural network for language s, $\mathbf{o}^{sL}$, is calculated as

$$\mathbf{o}^{sL} = \mathrm{softmax}(\mathbf{W}^{sL}\mathbf{o}^{L-1} + \mathbf{b}^{sL}) \tag{8}$$

where L is the output layer, $\mathbf{W}^{sL}$ and $\mathbf{b}^{sL}$ are the language-specific output weight and bias for language s. This architecture is similar to the multi-task multilingual training where the output layer consists of separate activations for each language, but it models the shared universal phone set. This approach will be investigated in end-to-end LF-MMI training in order to mitigate the side-effect from context dependent modelling and improve the multilingual models.

4. EXPERIMENTS

4.1. GlobalPhone Database

Experiments are reported on GlobalPhone [21]. We used the French (FR), German (GE), Portuguese (PO), Russian (RU) and Spanish (SP) datasets from the GlobalPhone corpus. 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[1]. The detailed statistics for each of the languages is shown in Table 1.

[1] 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.*

Language	Vocab	PPL	#Phones	Train	Dev	Eval
FR	65k	324	38	22.7	2.1	2.0
GE	38k	672	41	14.9	2.0	1.5
PO	62k	58	45	22.7	1.6	1.8
RU	293k	1310	48	21.1	2.7	2.4
SP	19k	154	40	17.6	2.0	1.7

4.2. Setup

We used 40-dimensional MFCC 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 unless otherwise stated. All the monolingual phones were mapped to IPA symbols and we merged the phonemes from FR, GE, PO, RU and SP to create the universal phone set for multilingual training.

The multilingual CTC model has 4 layers of Bidirectional Long Short-Term Memory (BLSTM), with 320 cells in each layer and direction. All the weights in the models were randomly initialized and were trained using stochastic gradient descent with momentum. A learning rate of 0.00004 was used and early stopping on the validation set was applied to select the best model. Dropout was applied as first proposed in [22]. The dropout rate was set to 0.2. For end-to-end LF-MMI training, 8 layers of Time Delay Neural Network (TDNN) was used, with 450 nodes in each layer. The network parameters are initialized randomly to have zero mean and a small variance. All CTC models were trained based on the EESEN implementation [23] and end-to-end LF-MMI systems were built using the Kaldi [24].

4.3. Results

4.3.1. Comparison Between CTC and End-to-end LF-MMI

Previous research has shown that the end-to-end LF-MMI training outperforms the CTC-based model on a fairly big dataset (Switchboard which has roughly 300 hours data). In this section, these two approaches are compared using much less data. Only monolingual data was used to train each model. Both monophone modelling and full biphone modelling were investigated.

Results are shown in Table2; it is clear that the end-to-end LF-MMI training significantly outperforms CTC training in low-resourced scenarios. It implies that LF-MMI training is less sensitive to the amount of training data. In addition, biphone modelling improves the performance in all the monolingual modeling cases.

Table 2. *Comparison between CTC training and end-to-end LF-MMI for monolingual low-resourced ASR in WER(%).*

system	FR	GE	PO	RU	SP
monophn CTC	26.9	24.3	21.0	32.7	11.7
biphn CTC	26.1	24.1	20.8	32.0	10.9
monophn LF-MMI	23.6	18.7	18.6	26.6	9.3
biphn LF-MMI	23.5	17.0	18.2	25.8	8.5

4.3.2. Multilingual Training

Multilingual training has been proved to be effective in traditional DNN-HMM training and CTC training. We further investigated multilingual training in the end-to-end LF-MMI framework. For multilingual biphone modelling, the pruned biphone targets were used as described in Section 3.2. The number of biphone targets was reduced from 23980 to 13776. The models were trained using data from all the 5 languages.

Table 3. *Comparison between multilingual CTC training and end-to-end LF-MMI in WER(%).*

system	FR	GE	PO	RU	SP
ML monophn CTC	24.9	23.6	19.6	31.6	10.7
ML monophn LF-MMI	23.2	15.4	17.0	24.9	7.9
ML biphn LF-MMI	23.2	16.0	17.9	25.1	7.7

Comparing Table 3 and Table 2, multilingual training yields significant improvement over monolingual training for both monophone and biphone-based LF-MMI. They both significantly outperform multilingual CTC training. However, different from the monolingual cases, the multilingual biphone LF-MMI performs worse than multilingual monophone model in most of the tested languages. We hypothesize that biphone targets cover more variabilities compared to the corresponding monophone, especially when they are shared by mutiliple languages. As reported in [20], language-specific characteristics cannot be well modeled by an IPA-based universal network. Language adaptive training (LAT) could be a solution to better model these variabilities across languages.

4.3.3. Language Adaptive Training

In order to test our hypothesis, language adaptive training was applied in both monophone and biphone based end-to-end LF-MMI training. The last two layers were trained to be language-specific. From Table 4, we can find that language adaptive training improves the performance for both cases and the biphone model benefits more from it. LAT helps further exploit the advantages of context-dependent modelling.

We further analyze the phoneme error rate with respect to the phonemes shared by multiple languages and the unique phonemes that only appear in one language. The reference phoneme sequences were extracted from the alignment and the hypotheses were generated from the best path of the decoding lattice. The analysis was conducted on the development sets of the 5 languages using monophone and biphone based models, as shown in Figure 1.

Table 4. *Comparison of multilingual end-to-end LF-MMI w/o LAT in WER(%).*

system	FR	GE	PO	RU	SP
ML monophn LF-MMI	23.2	15.4	17.0	24.9	7.9
+LAT	23.0	15.2	16.7	24.6	7.5
ML biphn LF-MMI	23.2	16.0	17.9	25.1	7.7
+LAT	22.7	14.8	16.6	24.1	7.3

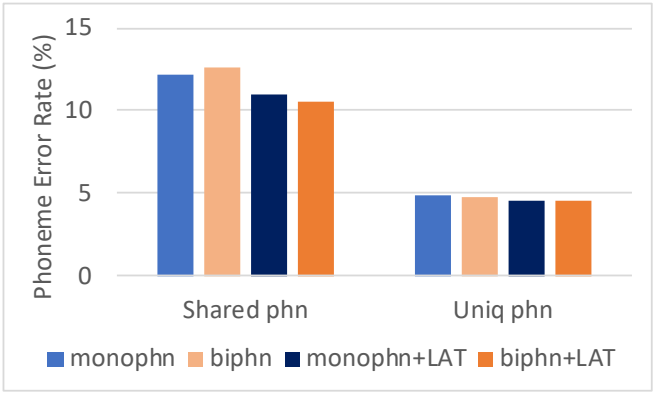

Fig. 1. PERs (%) comparison with or without Language Adaptive Training.

It is clear that biphone-based multilingual model indeed performs slightly worse than monophone-based model on the shared phones. However, it benefits more from the language adaptive training. Meanwhile, none of the models show much difference on the unique phonemes. This explains the above observations.

5. CONCLUSION

It was demonstrated that a universal phoneme-based multilingual TDNN trained with an end-to-end LF-MMI objective outperforms CTC training by a significant margin when training data is limited. Directly modelling biphones does not give any improvement over monophone modelling in multilingual training because of the increasing variations in context-dependent modelling. Language adaptive training can help break this bottleneck. The phoneme-based end-to-end LF-MMI training can be a better candidate also for cross-lingual adaptation, since the output layer can be easily extended to new phonemes as in CTC-based models. We leave this work as our future research.

6. ACKNOWLEDGEMENT

This work has been conducted with the support of the European Community H2020 Research and Innovation Action-funding, under “Scalable Understanding of Multilingual Media” (SUMMA) project No. 688139.

7. REFERENCES

[1] Ngoc Thang Vu and Tanja Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families," in *Proceedings of Interspeech*, 2013.

[2] Zoltán Tüske, Joel Pinto, Daniel Willett, and Ralf Schlüter, "Investigation on cross-and multilingual MLP features under matched and mismatched acoustical conditions," in *Proceedings ICASSP*, 2013.

[3] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings ICASSP*, 2013.

[4] Karel Veselỳ, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, 2012.

[5] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proceedings ICASSP*, 2014.

[6] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Proceedings ICASSP*, 2013.

[7] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, M. Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings ICASSP*, 2013.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006.

[9] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings ICASSP*, 2016.

[11] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proceedings of Interspeech*, 2018.

[12] Sibo Tong, Philip N Garner, and Hervé Bourlard, "Cross-lingual adaptation of a CTC-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[13] Sibo Tong, Philip N Garner, and Hervé Bourlard, "Fast language adaptation using phonological information," in *Proceedings of Interspeech*, 2018.

[14] Jeff Ma, Francis Keith, Tim Ng, Man-hung Siu, and Owen Kimball, "Improving deliverable speech-to-text systems with multilingual knowledge transfer," in *Proceedings of Interspeech*, 2017.

[15] Bhargav Pulugundla, Murali Karthick Baskar, Santosh Kesiraju, Ekaterina Egorova, Martin Karafiát, Lukáš Burget, and Jan Černockỳ, "BUT system for low resource indian language ASR," in *Proceedings of Interspeech*, 2018.

[16] Albert Zeyer, Eugen Beck, Ralf Schlüter, and Hermann Ney, "CTC in the context of generalized full-sum HMM training," in *Proceedings of Interspeech*, 2017.

[17] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI.," in *Proceedings of Interspeech*, 2016.

[18] Suyoun Kim and Michael L Seltzer, "Towards language-universal end-to-end speech recognition," *arXiv preprint arXiv:1711.02207*, 2017.

[19] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," *arXiv preprint arXiv:1711.01694*, 2017.

[20] Sibo Tong, Philip N Garner, and Hervé Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *Proceedings of Interspeech*, 2017.

[21] Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe, "GlobalPhone: A multilingual text & speech database in 20 languages," in *Proceedings ICASSP*, 2013.

[22] Jayadev Billa, "Improving LSTM-CTC based ASR performance in domains with limited training data," *arXiv preprint arXiv:1707.00722*, 2017.

[23] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.