IDIAP RESEARCH REPORT

LATTICE-FREE MAXIMUM MUTUAL INFORMATION TRAINING OF MULTILINGUAL SPEECH RECOGNITION SYSTEM

Srikanth Madikeri Banriskhem Khonglah
Sibo Tong Petr Motlicek Hervé Bourlard
Daniel Povey[a]

Idiap-RR-28-2020

NOVEMBER 2020

[a]Xiaomi Corporation, Beijing, China

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

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

Srikanth Madikeri[1], *Banriskhem K. Khonglah*[1], *Sibo Tong* [1,2], *Petr Motlicek*[1], *Hervé Bourlard*[1,2], *Daniel Povey*[3]

[1]Idiap Research Institute, Martigny, Switzerland

[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[3]Xiaomi Technology, 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 in the cost function. 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-resourced languages, training multilingual systems is an effective way to compensate for limited amount of data [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 [7, 8, 9]. Alternately, the models can be adapted to the target language [10, 11, 12, 13, 14, 15]. The multilingual models can either share the output layer or have separate output layers (one for each language) [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.

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.

In this work, we focus on sequence-discriminative training of multilingual AM with the Lattice-Free Maximum Mutual Information (LF-MMI) framework [16]. LF-MMI training has been shown to have superior performance compared to the conventional cross-entropy (CE) training of DNNs [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 [19]. The latter enforces the discriminative property in the training shown to be useful for training AM [20, 21].

Given the advantages of both multilingual and LF-MMI training procedures, it is natural to combine them to obtain better performance. In [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 [24][1]. 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.

[1]egs/babel_multilang/s5d/local/chain2/run_tdnn.sh

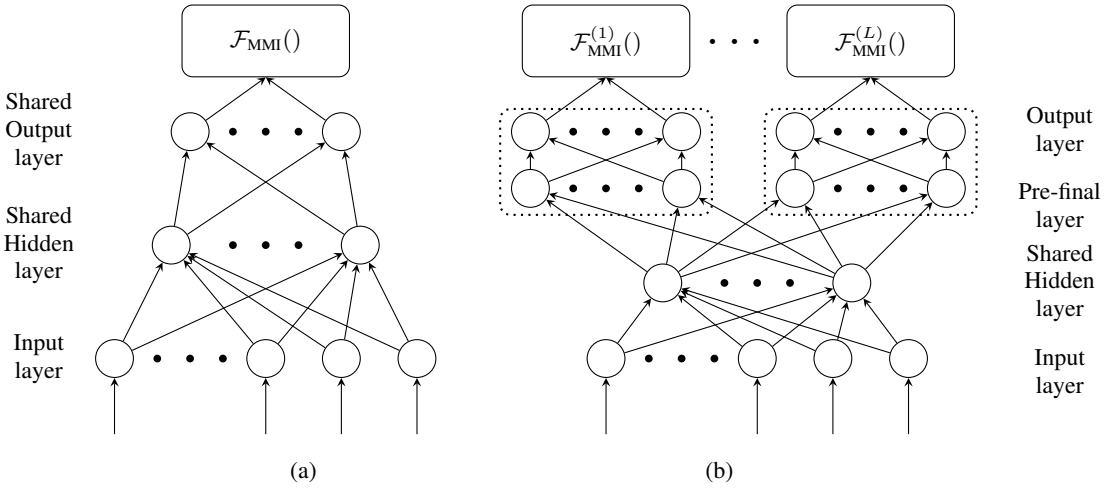

Figure 1: *(a) Multilingual LF-MMI system with shared output layer. The objective function $\mathcal{F}_{MMI}$ is computed with either a language-independent or language-dependent denominator graphs. (b) Proposed LF-MMI system with language-dependent objective functions. Both systems are shown for a simple feedforward architecture.*

2. LF-MMI

In LF-MMI, the MMI objective function is used as the cost function to train the AM [16]. The cost function is given as follows:

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

where $\mathbf{x}^{(u)}$ is the input sequence,

u is an utterance,

U is the set of all training utterances,

$\mathcal{M}_{\mathbf{w}(u)}$ corresponds to the numerator graph specific to a word sequence in transcription,

$\mathcal{M}_{\text{den}}$ is the denominator graph modelling all possible word sequences and

$\boldsymbol{\theta}$ is the model parameter.

The numerator can be computed either using alignments from another acoustic model, or in a completely end-to-end fashion [17]. In this work, we always use alignments from a monolingual HMM/GMM model.

The standard implementation of LF-MMI makes several simplifications to the conventional AM training of DNNs. First, the HMM topology is modified to a 2-state HMM so that the final state can be reached in one frame. Next, frame-dropping is employed during training so that only 1 in 3 frames is required during decoding. Finally, the segment length of an utterance during training is limited.

The derivatives of the two quantities–numerator and denominator–in Equation 1 are computed using two graphs. 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 by context-dependent states. Following the notation in [18], if ${}^{\text{NUM}}\gamma_t^{(u)}(s)$ is the posterior from the numerator at time t for state s and ${}^{\text{DEN}}\gamma_t^{(u)}(s)$ is that from the denominator, the gradient is given by:

$$\frac{\partial \mathcal{F}_{\text{MMI}}}{\partial y_t^{(u)}(s)} = {}^{\text{NUM}}\gamma_t^{(u)}(s) - {}^{\text{DEN}}\gamma_t^{(u)}(s), \qquad (2)$$

where $y_t^{(u)}(s)$ is the network output for state s at time t given input utterance u

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 $\mathcal{F}^{(1)}_{\text{MMI}}, \ldots, \mathcal{F}^{(L)}_{\text{MMI}}$ 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 functions, where L is the number of languages, computed independent of each other depending only on the language of the utterance:

$$\mathcal{F}_{\text{MMI}}^{(\ell)} = \sum_{u=1}^{U_\ell} \log \frac{p\left(\mathbf{x}^{(u)}|\mathcal{M}_{\mathbf{w}(u)}^{\ell}, \boldsymbol{\theta}\right) p(\mathbf{w}(u))}{p\left(\mathbf{x}^{(u)}|\mathcal{M}_{\text{den}}^{\ell}, \boldsymbol{\theta}\right)}, \quad (3)$$

where U_ℓ is the number of utterances in the minibatch for language ℓ, $\boldsymbol{\theta}$ contains the shared and language-dependent parameters, $\mathcal{M}_{\mathbf{w}(u)}^{\ell}$ and $\mathcal{M}_{\text{den}}^{\ell}$ 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 backpropogation, the shared parameters are then updated. The overall cost-function is the weighted sum of all language-dependent cost-functions:

$$\mathcal{F}_{\text{MMI}} = \sum_{\ell=1}^{L} \alpha_\ell F_{MMI}^{\ell}, \quad (4)$$

where α_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 multi-task 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.

[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.*

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

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)*

Language	Vocabulary	PPL	Train (h)	Eval (h)
Tagalog	22k	148	84.5	10.7
Swahili	25k	357	38.0	9.3
Turkish	41k	396	77.2	9.8
Zulu	56k	719	56.7	9.2

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].

We compare the monolingual systems to three multilingual systems: (1) single-task system trained with language independent denominator, (2) single-task system trained with language dependent denominator, and (3) multitask system trained with language dependent denominator. For the single-task systems, we concatenate the phonemes from the five languages to create the universal phone set for multilingual training. We did not use IPA-based phone set as in [23] because we found that the concatenated phone set performs better in preliminary experiments.

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).

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

System	FR	GE	PO	RU	SP
Monolingual LF-MMI	20.4	12.7	15.2	24.6	7.1
Single-task multilingual system					
Language independent	21.3	12.5	**14.9**	**22.1**	6.6
Language dependent	21.3	12.4	15.0	**22.1**	6.6
Multitask multilingual system					
5 languages	20.7	**11.7**	15.1	**22.1**	**6.5**

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

Category	Languages
Target languages & 4 Language Training	Tagalog, Swahili, Zulu, Turkish
14 Language Training	Tagalog, Swahili, Zulu, Turkish, Assamese, Bengali, Cantonese, Haitian, Kazhak, Kurmanji, Tamil, Telugu, Tok, Vietnamese

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 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.

4.4. BABEL results

The results on target languages from BABEL are presented in Table 5. The performance of the monolingual LF-MMI models are already better compared to those presented in literature, thus forming a strong baseline. Next, we compare the monolingual models to three multilingual models trained with the 4 language setup: (1) single-task system trained with language independent denominator, (2) single-task system trained with language dependent denominator, and (3) multitask system trained with language dependent denominator. The results show that in conditions with high acoustic variability, as in the case of BABEL data-sets, multilingual training brings considerable benefits. The multilingual systems show improvements over the monolingual systems for all languages. This clearly demonstrates the benefit of multilingual LF-MMI training for low-resource languages. Both single-task and multitask setups outperform the monolingual baseline, with relative improvements ranging from 2% to 8.8% for the former and 3% to 8% for the latter. In the single-task setup, as in the case of Globalphone, language-dependent denominator provides only marginal gains over language-independent denominator. Overall, the benefits obtained are dependent on the language, but no significant loss is observed by choosing one technique for multilingual training over the other for majority of the languages (Zulu being the exception).

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)*

System	TGL	SWA	TUR	ZUL
Monolingual LF-MMI	45.3	38.7	47.2	53.5
Single-task multilingual system				
Language independent	44.4	35.5	43.4	52.4
Language dependent	44.4	**35.4**	**43.0**	51.9
Multitask multilingual system				
4 languages	**43.9**	35.6	43.5	**51.0**
14 languages	**42.2**	**33.6**	43.9	**50.8**

To demonstrate the scalability of the multitask system, we also train an AM with 14 languages (final row in Table 5; the 14 languages are in Table 4). Compared with the 4 languages system, the 14 language system improves on 3 out of 4 languages. Relative improvements range from 0.4% (ZUL) to 5.6% (SWA) suggesting that adding more languages to the AM training can be beneficial without any additional cost during decoding. In addition, compared to the monolingual baseline relative improvement of up to 13.2% (SWA) is obtained.

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. References

[1] N. T. Vu and T. Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families." in *Interspeech*, 2013, pp. 515–519.

[2] L. Burget *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4334–4337.

[3] M. Karafiát *et al.*, "Multilingual blstm and speaker-specific vector adaptation in 2016 but babel system," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 637–643.

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[5] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech communication*, vol. 56, pp. 142–151, 2014.

[6] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7639–7643.

[7] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 336–341.

[8] F. Grézl, M. Karafiát, and K. Veselỳ, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7654–7658.

[9] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7349–7353.

[10] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7319–7323.

[11] S. Tong, P. N. Garner, and H. Bourlard, "Cross-lingual adaptation of a ctc-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[13] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 2322 – 2326.

[14] P. Motlicek, D. Imseng, B. Potard, P. N. Garner, and I. Himawan, "Exploiting foreign resources for dnn-based asr," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 2015:17, Jun. 2015.

[15] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proceedings of Interspeech 2017*, Aug. 2017, pp. 2406–2410.

[16] D. Povey *et al.*, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[17] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. Interspeech*, 2018, pp. 12–16.

[18] H. Hadian *et al.*, "Flat-start single-stage discriminatively trained hmm-based models for asr," *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1949–1961, 2018.

[19] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2005.

[20] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, vol. 2013, 2013, pp. 2345–2349.

[21] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3761–3764.

[22] F. Keith, W. Hartmann, M.-H. Siu, J. Ma, and O. Kimball, "Optimizing multilingual knowledge transfer for time-delay neural networks with low-rank factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4924–4928.

[23] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of multilingual asr using end-to-end lf-mmi," *In Proc. of ICASSP 2019*, pp. 6061–6065, 2019.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[25] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.

[26] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.