# Multitask adaptation with Lattice-Free MMI for multi-genre speech recognition of low resource languages

*Srikanth Madikeri*[1], *Petr Motlicek*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
srikanth.madikeri, petr.motlicek, herve.bourlard@idiap.ch

## Abstract

In this paper, we develop Automatic Speech Recognition (ASR) systems for multi-genre speech recognition of low-resource languages where training data is predominantly conversational speech but test data can be in one of the following genres: news broadcast, topical broadcast and conversational speech. ASR for low-resource languages is often developed by adapting a pre-trained model to a target language. When training data is predominantly from one genre and limited, the system's performance for other genres suffer. To handle such out-of-domain scenarios, we employ multitask adaptation by using auxiliary conversational speech data from other languages in addition to the target-language data. We aim to (1) improve adaptation through implicit data augmentation by adding other languages as auxiliary tasks, and (2) prevent the acoustic model from overfitting to the dominant genre in the training set. Pre-trained parameters are obtained from a multilingual model trained with data from 18 languages using the Lattice-Free Maximum Mutual Information (LF-MMI) criterion. The adaptation is performed with the LF-MMI criterion. We present results on MATERIAL datasets for three languages: Kazakh and Farsi and Pashto.

**Index Terms**: Lattice Free MMI, low-resource speech recognition, multitask learning

## 1. Introduction

In the MATERIAL (Machine Translation for English Retrieval of Information in Any Language[1]) program, ASR systems for low-resource languages are trained on predominantly conversational speech, but tested on speech from multiple genres: conversational speech (CS), news broadcast (NB) and topical broadcast (TB). For such tasks, an ASR that generalizes better across multiple genres despite the constraints imposed on the training data is desirable. Owing to the low-resource nature of the target languages, a common approach is to adapt a pre-trained model to the target language [1, 2]. Multilingual modelling is a common technique used to boost training resources for the acoustic model [3, 4, 5, 6]. In the Babel program [7], multilingual models were trained using data from all languages in the program [8, 9], which proved to be effective on both seen and unseen languages in training.

In [10, 11], adaptation of pre-trained Lattice Free-Maximum Mutual Information Criterion (LF-MMI) models was shown to be effective for ASR on out-of-domain data. In [12], multilingual models trained with the LF-MMI were shown to outperform monolingual models on both Babel and Globalphone datasets. In this paper, we show the effectiveness of adapting such multilingual LF-MMI models to MATERIAL's multi-genre test condition. Compared to training monolingual models with LF-MMI, adaptation of multilingual LF-MMI models perform significantly better across all genres. Similar to [11], we adapt the multilingual model by adding new language-specific output layers, even in the case where languages were seen during multilingual training.

Given a target-language, ASR can be trained by simply adapting existing language specific layers in the model, or adding new layers to be trained during adaptation. In the latter case, typically the learning rate on the pre-trained layers is a fraction of the learning on newly added layers (e.g. one tenth of the learning rate of the new layers [10]). Since the amount of adaptation data is limited (few tens of hours of speech) and mostly from a single domain (CS), the model tends to adapt well towards the genre predominant in the training data.

Unlike CS, broadcast speech data for many languages are available in the open source domain. Thus, to improve the performance on broadcast data one can further perform semi-supervised training (SST) [13, 14, 15, 16, 17]. Moreover, as shown in [16], improving the seed model can provide a considerable boost to the final performance on broadcast data with SST. Thus, we propose a simple approach using multitask learning that can provide a better starting point for techniques such as SST. The goal of applying multitask learning for adaptation (or transfer learning in the case where the target language is unseen during multilingual training) is to use auxiliary tasks as competing objectives to boost the adapted model's out-of-domain performance. Given the success of multilingual LF-MMI training, we extend it to target language adaptation as well. In this case, we consider models pre-trained with multilingual LF-MMI with 18 languages. The model is adapted, also with the LF-MMI criterion, along with other languages that are not necessarily our target. We refer to this technique as multitask adaptation (MTA), while the conventional adaptation of pre-trained models is referred to as Single Task Adaptation (STA). On MATERIAL datasets, we show that by replacing STA with MTA, one can achieve relative improvements in Word Error Rate (WER) of up to 7.1%. We will release the MTA adaptation code as part of the Babel multilingual recipe in Pkwrap [18] to adapt both Kaldi and Pytorch [19] acoustic models trained with LF-MMI [2].

The rest of the paper is organized as follows: in Sections 2 and 3, multilingual LF-MMI training and our multitask adaptation method are described, respectively. In Section 4, we detail the proposed approach of multi-task adaptation. In Section 5, experimental details and results are presented.

[1]https://www.iarpa.gov/index.php/research-programs/material

[2]https://github.com/idiap/pkwrap/tree/master/egs/multilang/babel/

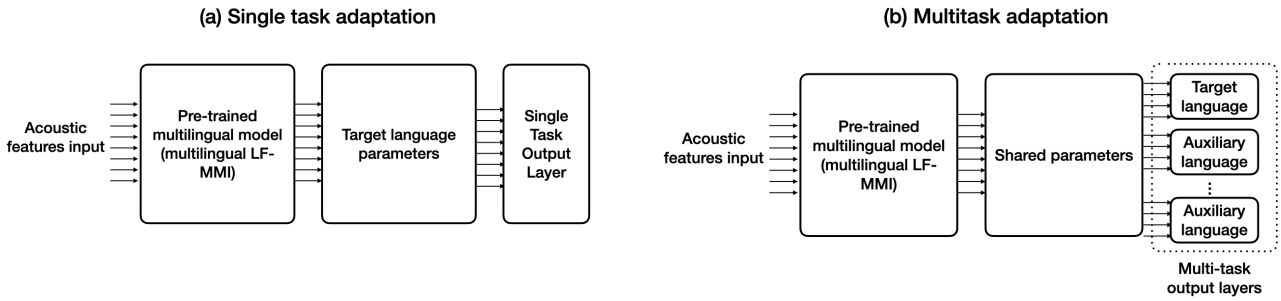


Figure 1: *(a) illustration of typical adaptation of pre-trained model to a target-language. (b) illustration of the proposed multitask adaptation with target language as one of the tasks. The target language shares parameters with auxiliary tasks (other languages used during adaptation).*

## 2. Multilingual LF-MMI

In [12], a multitask setup to train multilingual acoustic models with LF-MMI was introduced. The LF-MMI criterion provides state-of-the-art performance for hybrid ASR systems. LF-MMI provides a sequence discriminative training criterion, wherein each sequence (typically, an utterance of speech) is evaluated by two values: the numerator which computes the probability of the observation given the groundtruth, and the denominator which computes the probability over all possible sequences. The latter is computed with a graph, referred to as the denominator graph, trained from a phone Language Model (LM) [20]. The phone LM is trained from transcripts in the training data. In multilingual LF-MMI, the acoustic model shares parameters across languages, and there is one output layer for each language in the training dataset. Each language has its own denominator graph during training.

The performance of multilingual models on the Babel datasets is well established with standard Time Delay Neural Networks (TDNN) [21]. In this paper, we improve the model capacity of the AM by using the CNN-TDNN-F architecture (Convolutional Neural Networks and Factorized TDNNs) trained with 18 languages obtained from Babel and MATERIAL datasets (as opposed to only 14 in our previous work), thereby learning better representations suitable for cross-lingual learning [22, 23]. The list of datasets used for trained are given in Table 1. Note that we only refer to the multitask version of multilingual training in this paper, where each language in training has a separate output layer.

We apply transfer learning on this multilingual model to languages recently considered in the MATERIAL program: Pashto, Farsi and Kazakh. Out of the three, two languages, Pashto and Kazakh, overlap with the 18 languages used for multilingual training. Farsi is treated as an unseen language. The adaptation is carried out in a fashion similar to [10]. We do not freeze all the layers in the multilingual model, but fix a learning rate factor on the pre-trained layers. To adapt to each language, a learning rate factor of 0.1 was used. In addition to the pre-trained layers we also add additional target language-specific layers. To control the number of model parameters, we use TDNN-F layers [24]. The LF-MMI criterion is used for adaptation.

## 3. Multitask adaptation

In this section, we describe the proposed multitask approach. To motivate our approach we provide the following reasoning: in order to improve the AM for low-resource languages, multilingual modelling is often considered useful. Similarly, when adapting a well-trained acoustic model to a target language, one can employ a similar strategy by adapting multiple languages at the same time despite our interest being in only one of the languages. As mentioned earlier, we refer to this type of adaptation as Multitask adaptation (MTA). To contrast with MTA, we will refer to the conventional adaptation of pre-trained models to a target language as Single Task Adaptation (STA). In Natural Language Processing tasks, where using pre-trained models is quite common, MTA of pre-trained models has been shown to be effective [25]. Figure 1 illustrates the difference between STA and MTA.

Multitask learning [26, 27] has several well-documented advantages. Two important advantages that we consider here are implicit data augmentation and ability to reduce the risk of overfitting. When adapting pre-trained models to low-resource languages, we observed that despite heavy regularization through high dropout rates, the model performance saturates. To avoid such saturation we use the regularizing effect of adding new languages. Multitask learning for regularization has already been applied in different contexts. In LF-MMI training, it is common to use cross-entropy objective function as an auxiliary objective. In end-to-end ASR training, using multiple objective functions has been shown to be useful [28].

In addition, the presence of more data from different languages is well-known to improve speech models [29, 30, 31]. Thus, we hypothesize that adapting a pre-trained model to multiple languages instead of just the target language can be more beneficial to the performance on out-of-domain data. In this work, we consider four languages for MTA: Kazakh, Farsi, Pashto and Turkish. The first three are target languages, and Turkish is included due to its linguistic proximity to Kazakh (among the Babel datasets used in this work). In order to balance the trade-off between the adaptation speed and multi-task adaptation benefits, we do not consider more than four languages.

## 4. Experiments

We first evaluate the performance of the improved multilingual model on four languages from Babel: Tagalog (TGL), Swahili (SWA), Zulu (ZUL) and Turkish (TUR). The evaluation setup for Babel is the same as [12]. Then, we report the results on three languages in the MATERIAL program: Farsi, Kazakh and Pashto.

| | | | |
|---|---|---|---|
| Assamese | Bengali | Cantonese | Haitian |
| Kazhak | Kurmanji Kurdish | Lao | Lithuanian |
| Pashto | Somali* | Swahili | Tagalog |
| Tamil | Telugu | Tok Pisin | Turkish |
| Vietnamese | Zulu | | |

Table 1: *Babel [7] and MATERIAL (marked with *) datasets used for multilingual training. The language names are sorted in alphabetical order.*

| Layer | Parameter |
|---|---|
| CNN-1 | 64 filters |
| CNN-2 | 64 filters |
| CNN-3 | 128 filters + height subsampling |
| CNN-4 | 128 filters |
| CNN-5 | 256 filters + height subsampling |
| CNN-6 | 256 filters |
| TDNN-F | 1536 dim, 256 dim BN |
| TDNN-F x 7 | 1563 dim + 0.66 bypass scale |
| Bottleneck layer | 512 dimension |

Table 2: *Description of the architecture of the multilingual CNN-TDNN-F model. The architecture is a modification of a similar model found in standard Kaldi recipes (`egs/librispeech/s5/local/chain/tuning/run_cnn_tdnn_1a.sh`). (dim: dimension, BN: bottleneck)*

### 4.1. Model training

The multilingual model was trained with the 18 languages given in Table 1. For all Babel datasets, only conversational speech data was used for training. We trained a 14-layer CNN-TDNN-F (Convolutional Neural Network followed by Factorized Time-delay Neural Networks [24]). The model architecture is given in Table 2. We used hybrid LF-MMI to train the model, with a weight of 1/18 for each language. The model takes as input 40 dimensional MFCC features and online i-vectors ([32, 33]). Three-fold speed-perturbation was applied to the training data.

To generate alignments for training, a HMM/GMM system was trained with PLP+pitch (a concatenation of Perceptual Linear Prediction and pitch) features using the standard recipe for Babel datasets in Kaldi [34]. The lexicon provided with the dataset was used. The alignments generated were used to create supervision lattices for LF-MMI training. The acoustic model was trained for 6 epochs with an exponentially decaying learning rate schedule with an initial learning rate of 0.001 and final learning rate of 0.0001. A dropout schedule with the following parameters was used: from 20% to 50% of the iterations, the dropout was increased from 0.0 to 0.25, and then was gradually decreased to 0.0 for the rest of the iterations. A continuous version of a dropout was used [34]. We used Kaldi for all our experiments.

### 4.2. Performance on Babel

The performance of the multilingual model on four languages is presented in Table 3. WERs are reported on dev10h test set. We also refer to performance reported in [35] to compare with our baseline monolingual systems. As reported in [12], the multilingual model trained with 14 languages is significantly better than the monolingual LF-MMI system. Relative improvements of up to 13.6% (SWA) was achieved. From the results with the CNN-TDNN-F model, it is clear that the multilingual training can further benefit with increased model capacity. The CNN-TDNN-F model improves further by 6.6% for TGL, 7.1% for SWA.

| System | TGL | SWA | TUR | ZUL |
|---|---|---|---|---|
| Monolingual TDNN [12] | 45.3 | 38.7 | 47.2 | 53.5 |
| BLSTM [35] | 46.3 | 38.3 | - | 61.1 |
| Multilingual models | | | | |
| TDNN (14 languages) [12] | 42.2 | 33.6 | 43.9 | 50.8 |
| CNN-TDNN-F (18 languages) | 39.4 | 31.2 | 40.8 | 48.5 |

Table 3: *Comparison of performance of multilingual LF-MMI models on four languages in the Babel dataset. Word Error Rates (WER) on dev10h are reported. We also compare our results with [35] as reference to other multilingual models with similar datasets.*

| Parameter | Pashto | Kazakh | Farsi |
|---|---|---|---|
| Training data (h) | 78.4 | 49.8 | 36.3 |
| Test data (CS, NB, TB) (h) | 16.4 | 11.2 | 9.5 |
| Vocabulary | 239k | 580k | 1.7M |
| LM (words) | 816k | 184M | 1.3B |
| LM Perplexity (3-gram) | 560 | 789 | 786 |

Table 4: *Statistics of the MATERIAL test sets for Pashto, Kazakh and Farsi. Train and test data duration are computed after segmentation. The segmentation is taken from groundtruth. LM perplexities are calculated with the LM trained on all text available for the language and evaluated on only broadcast data transcripts.*

### 4.3. MATERIAL datasets

We consider three MATERIAL datasets: Kazakh, Pashto and Farsi. The first two languages are also part of the Babel datasets used for multilingual training while Farsi is an unseen language.

Language model for each dataset is trained as follows: for each language text obtained from web-crawl is available for language model. The web-crawl text is cleaned (punctuation and out-of-language word) and a 3-gram model is trained with SRILM [36] along with the training transcripts. We use Kneser-Ney smoothing with parameters 0, 1 and 2. This consistently gave us the best trade-off between language model perplexity and size. This language model is used for decoding NB and TB audio. For CS, we interpolate the LM with a 3-gram LM trained only with training transcripts. An interpolation weight of 0.9 on the latter is used [37]. The vocabulary for each language is chosen based on the web crawl text and training transcripts. While all words in the training transcripts are included, only words that appear at least 5 times in the web crawl are chosen as a part of the vocabulary. Graphemic lexicon was used for all the

| | System | *Seen* languages | | | | | | *Unseen* language | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Pashto | | | Kazakh | | | Farsi | | |
| | | CS | NB | TB | CS | NB | TB | CS | NB | TB |
| (a) | Monolingual TDNN-F | 47.2 | 47.0 | 54.8 | 44.3 | 29.4 | 36.2 | 50.7 | 56.6 | 49.7 |
| (b) | Monolingual CNN-TDNNF | 46.9 | 44.2 | 51.3 | 39.7 | 25.9 | 30.9 | 43.2 | 42.4 | 48.9 |
| (c) | STA | 41.9 | 43.6 | 48.1 | 39.2 | 23.4 | 26.6 | 37.0 | 36.6 | 41.1 |
| (d) | MTA | 41.8 | **40.5** | **45.4** | 38.9 | **21.9** | **25.4** | 36.9 | **35.3** | **40.1** |
| (e) | Fusion (c+d) | 40.8 | 40.7 | 45.2 | 37.6 | 21.6 | 24.7 | 35.3 | 33.8 | 38.6 |

Table 5: *Comparison of performance of adaptation with multilingual LF-MMI models to three MATERIAL datasets. Word Error Rates (WER) are reported. CS: Conversational speech, NB: News Broadcast, TB: Topical Broadcast, STA: Single task adaptation, MTA: Multitask adaptation*

three languages. All words in Kazakh were lower-cased. The statistics of training data is given in Table 4.

Two experiments were performed on the MATERIAL languages: (1) STA (adaptation of the multilingual CNN-TDNN-F model to the target language), and (2) MTA (multitask adaptation of the same pre-trained model to several target languages, simultaneously). We also used Babel Turkish as an additional language for MTA. The adaptation was carried out by setting a learning rate factor of 0.1 on the pre-trained layers. Additional 9 layers of TDNN-F was added to adapt to each target language. All but the first TDNN-F component had a context of 3. The first TDNN-F layer takes as input the output of the bottleneck layer of the multilingual model. The same network architecture was used for both STA and MTA. Each output layer in MTA had a learning rate factor of 0.25 (i.e. all languages were weighted equally). An exponentially decaying learning rate schedule was used with initial learning rate of 0.001 and final learning rate of 0.0005. A different dropout schedule was used during adaptation: dropout rate was kept to 0.0 for the first 5% of the iterations, then increased to 0.25 until 60% of the iterations, followed by reduction to 0.0 until the final iteration.

### 4.4. Performance on MATERIAL datasets

The results are presented in Table 5. First we compare the results of monolingual systems with systems adapted from the multilingual model. Considerable improvements are observed for all 3 languages. The benefits of adapting a multilingual model with STA is shown by relative improvements obtained up to 15.9% (Farsi, TB) compared to the best monolingual system. All systems performed the worst on the TB compared to other genres owing to the difficulty of the genre (mostly in terms of acoustic conditions and vocabulary). Adapting any of the three target languages provides significant performance boost for all genres.

With MTA, improvements in the broadcast genre (i.e. NB and TB) were observed for all languages. The results demonstrate that MTA can be beneficial compared to STA for out-of-domain data. Note that for both STA and MTA the same model configuration is used. Relative improvements ranging from 2.5% (TB in Farsi) to 7.1% (NB in Pashto) are observed for the broadcast genre. For in-domain data (CS), we only observed marginal gain in performance. However, to verify if the acoustic model trained with MTA is different to that obtained with STA, we performed a simple system fusion experiment. Improvements observed on 8 out of the 9 subsets suggest that MTA learns representations different to that learnt with STA. Even though the difference between STA and MTA performances are negligible for the CS genre, the fusion of the two systems provided relative improvements between 2.4% (Pashto) and 4.4% (Farsi). For NB in Pashto, there is a slight degradation in performance (from 40.5% to 40.7%) suggesting that the acoustic representation obtained with MTA can sometimes be considerably better for broadcast data than that obtained with STA.

## 5. Summary

We presented results on four Babel languages with multilingual LF-MMI training. We showed that multilingual LF-MMI scales well with increased model capacity, and with the number of languages used during training. We demonstrated the usefulness of such pre-trained models for multi-genre speech recognition on the MATERIAL dataset for three languages: Pashto, Kazakh and Farsi. Consistent improvements were obtained for both seen and unseen languages. To further improve the performance on broadcast data we proposed multitask adaptation. Relative improvements ranging between 2.5% and 7.1% were obtained compared to the conventional adaptation on news and topical broadcast.

## 6. Acknowledgements

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.

We thank Joel Barry (ISI, USC) for sharing the web-crawl text for the MATERIAL languages used in this paper.

## 7. References

[1] F. Grézl, M. Karafiát, and K. Veselỳ, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. of ICASSP 2014*, pp. 7654–7658.

[2] D. Bagchi and W. Hartmann, "Learning from the best: A teacher-student multilingual framework for low-resource languages," in *Proc. of ICASSP 2019*, pp. 6051–6055.

[3] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *In Proc. of Interspeech*, 2012, pp. 1191–1194.

[4] 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*, vol. 2015, no. 1, pp. 1–10, 2015.

[5] P. Motlicek, F. Valente, and P. N. Garner, "English spoken term detection in multilingual recordings," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] D. Imseng, P. Motlicek, P. N. Garner, and H. Bourlard, "Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 332–337.

[7] M. Harper, "The BABEL program and low resource speech technology," *In Proc. of Automatic Speech Recognition and Understanding*, 2013.

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

[9] K. M. Knill *et al.*, "Investigation of multilingual deep neural networks for spoken term detection," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 138–143.

[10] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for asr using lf-mmi trained neural networks," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2017, pp. 279–286.

[11] A. Vyas, S. Madikeri, and H. Bourlard, "Lattice-free MMI adaptation of self-supervised pretrained acoustic models," *In Proc. of ICASSP 2021*, pp. 6219–6223.

[12] S. Madikeri *et al.*, "Lattice-free maximum mutual information training of multilingual speech recognition systems," in *Proc. of Interspeech*, 2020, pp. 4746–4750.

[13] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. of ICASSP 2006*, pp. 1056–1059.

[14] K. Veselỳ, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 267–272.

[15] D. Imseng *et al.*, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proc. of ICASSP*, 2014, pp. 2322–2326.

[16] B. Khonglah *et al.*, "Incremental semi-supervised learning for multi-genre speech recognition," in *Proc. of ICASSP*, 2020, pp. 7419–7423.

[17] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4844–4848.

[18] S. Madikeri, S. Tong, J. Zuluaga-Gomez, A. Vyas, P. Motlicek, and H. Bourlard, "Pkwrap: a pytorch package for LF-MMI training of acoustic models," *arXiv preprint arXiv:2010.03466*, 2020.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[20] D. Povey *et al.*, "Purely sequence-trained neural networks for asr based on Lattice-Free MMI," in *Proc. of Interspeech*, 2016, pp. 2751–2755.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. of Interspeech*, 2015, pp. 3214–3218.

[22] I. Medennikov, Y. Y. Khokhlov, A. Romanenko, I. Sorokin, A. Mitrofanov, V. Bataev, A. Andrusenko, T. Prisyach, M. Korenevskaya, O. Petrov *et al.*, "The STC ASR system for the voices from a distance challenge 2019." in *In Proc. of INTERSPEECH*, 2019, pp. 2453–2457.

[23] M. Karafiát, M. K. Baskar, I. Szöke, H. K. Vydana, K. Veselỳ, J. Černockỳ *et al.*, "BUT Opensat 2019 speech recognition system," *arXiv preprint arXiv:2001.11360*, 2020.

[24] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Proc. of Interspeech*, 2018, pp. 3743–3747.

[25] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, 2019.

[26] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[27] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[28] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," *Proc. of Interspeech*, pp. 1–5, 2020.

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

[30] N. T. Vu *et al.*, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. of ICASSP*, 2014, pp. 7639–7643.

[31] 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.

[32] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[33] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. of ICASSP*, 2014, pp. 225–229.

[34] 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.

[35] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *Proc. of ICASSP*, 2019, pp. 6096–6100.

[36] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[37] E. Boschee *et al.*, "Saral: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 19–24.