Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6444–6449
Marseille, 11–16 May 2020

Automatic Speech Recognition for Uyghur through Multilingual Acoustic Modeling

Ayimunishagu Abulimiti, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany
{ay.abulimiti, tanja.schultz}@uni-bremen.de

Abstract

Low-resource languages suffer from lower performance of Automatic Speech Recognition (ASR) system due to the lack of data. As a common approach, multilingual training has been applied to achieve more context coverage and has shown better performance over the monolingual training (Heigold et al., 2013). However, the difference between the donor language and the target language may distort the acoustic model trained with multilingual data, especially when much larger amount of data from donor languages is used for training the models of low-resource language. This paper presents our effort towards improving the performance of ASR system for the under-resourced Uyghur language with multilingual acoustic training. For the developing of multilingual speech recognition system for Uyghur, we used Turkish as donor language, which we selected from GlobalPhone corpus as the most similar language to Uyghur. By generating subsets of Uyghur training data, we explored the performance of multilingual speech recognition systems trained with different sizes of Uyghur and Turkish data. The best speech recognition system for Uyghur is achieved by multilingual training using all Uyghur data (10 hours) and 17 hours of Turkish data and the WER is 19.17%, which corresponds to 4.95% relative improvement over monolingual training.

Keywords: ASR, Low- resource, Multilingual training, Agglutinative languages, GlobalPhone

1. Introduction

Speech recognition technology has gained dramatic improvement recently and has shown promising results on many tasks (Graves et al., 2006; Hinton et al., 2012; Chiu et al., 2018). However, data sparsity is still an issue for training more reliable acoustic and language models. Compared to resource-rich languages, low-resource languages suffer from higher Word Error Rate (WER) in speech recognition tasks. It has been long established that acoustic models trained across multiple languages can partly compensate this resources gap (Schultz and Waibel, 2001). In particular, data from resource-rich languages, which are similar to the target low-resource language, might be used in multilingual training with the aim to cover more contextual variations, thereby improving the performance of Automatic Speech Recognition (ASR) system for low-resource languages (Caruana, 1997; Heigold et al., 2013; Ghahremani et al., 2017; Huang et al., 2013; Sahraeian and Van Compernolle, 2016). However, the impurity of training data that comes from different sources (donor languages and target language), may hurt the acoustic model of target language and even the donor language (Vu and Schultz, 2013; Lin et al., 2009). Therefore, finding proper balance between the amount of data from target and donor language is one issue in developing multilingual ASR for low-resource language.

Uyghur language is an under-resourced language with about 11 million speakers, who are mainly located in western China and Central Asia. Uyghur belongs to the Turkic family. It is similar to Turkish with agglutinative morphology, object-verb constituent order and vowel/constant harmonic processes. The similarity of pronoun and numbers also contribute to the mutual intelligibility of these languages.

In this work, we explored the performance of different multilingual speech recognition systems trained with different sizes of data from low-resource language and donor language. We developed hybrid Hidden Markov Model (HMM) / Deep Neural Network (DNN) based monolingual and multilingual speech recognition systems for different data size of the target language, Uyghur. The goal of the study is to investigate the improvements of multilingual systems over monolingual systems when using different amount of Uyghur data in the training.

For developing multilingual speech recognition system for Uyghur, we used Turkish data from GlobalPhone corpus as donor language data, since both of the languages are close to each other in terms of linguistic structure.

Unlike multilingual DNN acoustic models that trained with shared hidden layers and have language specific softmax layers, our multilingual DNN acoustic models have shared hidden layers as well as softmax layers.

The paper is organized as follows: in the next section, the speech and text corpora of Uyghur will be described. In Section 3, monolingual and multilingual training experiments will be introduced. In Section 4, the results from our experiments will be discussed. The paper is concluded with remarks from multilingual training in Section 5.

2. Data

The Uyghur and Turkish text data used in this study were collected by applying the GlobalPhone corpus collection procedures as described in (Schultz, 2002). As of today, the Globalphone corpus comprises of more than 450 hours of high-quality clean speech recorded from more than 2000 native speakers reading newspaper articles available from the web (Schultz et al., 2013).

Using uniform GlobalPhone-style data has several benefits such as high chances of reliable alignments due to the planned speaking style (read) and high-quality signals. Furthermore, the same domain across all languages allows comparisons across vocabulary size, word usage, and statistical estimates of language complexity. However, it remains to be seen whether the results can be transferred to spontaneous speaking style.

2.1. Speech Corpus

The Uyghur data collection, partially funded by NSF (award 1519164), comprises of news articles read by 46 speakers. Each speaker read about 100 utterances resulting in a total of 4271 utterances. The speech was recorded with linear 16-bit Pulse Code Modulation (PCM) with 44.1 kHz sampling rate in clean surrounding conditions. The speech data was down-sampled to 16 kHz to keep the GlobalPhone speech data style. For the purpose of developing ASR for Uyghur language, 80% of all recordings were divided into training, 10% was divided into development and evaluation set with the constraint that no speaker and no utterance appears in more than one set. The details of the Uyghur data distribution is shown in Table 1. In multilingual training, we used the Turkish data from GlobalPhone corpus. In Table 2, the distribution of Turkish data is presented.

For further details of Turkish data refer to Schultz (2002).

	Training	Development	Evaluation
Speakers	37	4	5
Audio length	10:48	01:23	01:54
Utterances	3380	400	491
Word tokens	60084	7902	9741

Table 1: Uyghur speech database (audio length given as hours:min)

	Training	Development	Evaluation
Speakers	79	11	10
Audio length	13:12	1:58	1:53
Utterances	5482	734	731
Word tokens	87733	12381	12552

Table 2: Turkish speech database (audio length given as hours:min)

2.2. Text Corpus

There are three writing systems (Arabic, Latin and Cyrillic alphabet) in Uyghur. The newspaper articles, which correspond to audio, were written in Uyghur Arabic form. In Arabic written form, short vowels are not included in the writing form. As a consequence, it is not convenient to generate pronunciation dictionary based on grapheme-to-phoneme property. However, in Latin or Cyrillic written form, the pronunciation is very close to the written form, so that pronunciation of words can be generated base on grapheme-to-phoneme converter. For that reason, the written form in Latin is preferred in developing speech recognition system and the texts in Arabic written form were converted into Uyghur Latin. Since these newspaper articles were not sufficient for building an accurate language model, additional online broadcast texts were also used for language modelling. Therefore, the broadcast texts were also converted into Uyghur Latin and then normalized (special characters and punctuation were removed, numbers were converted into text). After these pre-processing, the broadcast texts contained about 10 million word tokens and 250k unigram words.

2.3. Pronounciation Dictionary

GlobalPhone dictionaries are based on International Phonetic Alphabet (IPA) scheme. In the first stage, the IPA for Uyghur was generated based on the Uyghur-Latin. Based on the IPA, the phones are mapped to the GlobalPhone phones, in order to make it available for the multilingual speech recognition development and keep the GlobalPhone dictionary style consistent.

We used 32 basic phonemes consisting of 8 vowels and 24 consonants for Uyghur language. The pronunciation dictionary consists of 49k words, which covers all words in training data and the selected words from broadcast texts (described in Section 2.2.). The Out Of Vocabulary (OOV) rate on evaluation set is 4.61%.

In GlobalPhone dictionary, Turkish dictionary contains 29 phonemes (8 vowels, 21 consonants) and 33.6k words. The OOV rate on evaluation set is 1.25%.

27 phonemes are included both in Uyghur and Turkish dictionary. Two phonemes *ğ (IPA: j)* and *ı (IPA:* ı*)* of Turkish are not included in Uyghur and five phonemes *ë (IPA: /e/)*, *ng (IPA:* ŋ*)*, *q (IPA: q)*, *h (IPA:* χ*)*, *gh (IPA:* ʁ*)* of Uyghur are not included in Turkish.

2.4. Similar Property of Uyghur and Turkish

When selecting a similar language to Uyghur from GlobalPhone for multilingual training experiments, the language family groups are considered. In addition, Type-to-Token Ratio (TTR), which indicates the lexical richness of a language is also taken into account. The TTR value of Uyghur and Turkish is close and 14.35% and 15.28% (Tachbelie et al., 2020), respectively.

As showed in Figure 1, we also analyzed the distribution of number of phoneme per word in the lexicon of both languages, which can reflect the aggultinative morphology of two languages (Carki et al., 2000). The frequency distribution over phoneme length per word in the lexicon of both languages are similar and in both of the languages, the most frequent phoneme length is 8.

Regarding these factors, we selected Turkish as donor language to Uyghur and used Turkish data for multilingual training for speech recognition system for Uyghur. The details of Turkish data can be found in (Schultz, 2002).

3. Experiments

3.1. Acoustic Model

In speech recognition tasks for low-resource languages, the main purpose of the multilingual training is to achieve more context coverage. However, the difference between the donor and target language used in training may hurt the acoustic model trained with multilingual data, especially when much larger amount of donor language data is used for the multilingual training. To investigate how the proportion of data from low-resource and resource-rich languages affect the performance of ASR system for low-resource language, we conducted two sets of experiments, monolingual and multilingual training, each with five different training

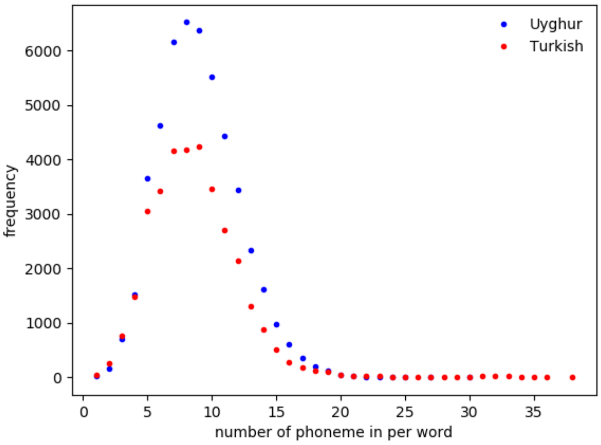

Figure 1: Phoneme length per word in lexicon of Uyghur and Turkish

sets. To simulate different levels of low-resource conditions, we randomly selected Uyghur data from the whole Uyghur training data with varying sizes. The selected Uyghur speech data duration ranges from 2 to 10 hours with 2-hour increment. Thus, five data sets are generated and in every set, all the speakers are included. These five sets of Uyghur data are combined with whole Turkish training data to perform multilingual training. Monolingual systems were built by only using the five selected Uyghur subset data.

In both monolingual and multilingual experiments, hybrid HMM/DNN systems were developed. All recognition systems were build using the open-source Kaldi ASR toolkit (Povey et al., 2011). First, context dependent HMM/Gaussian Mixture Model (GMM) based speech recognition systems were build using 39 dimensional stacked Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) for the alignments of DNN training. We applied Speaker based Cepstral Mean and Variance Normalization (CMVN) with context size of 7 frames. And then Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transform (MLLT) model was generated. At the end, Speaker Adaptive Training (SAT) was conducted using an affine transform, feature space Maximum Likelihood Linear Regression (fMLLR). In all experiments, fMMLR models showed the best results among monophone, triphone and LDA+MLLT models. Thus, in every experiment, the alignments from fMLLR are used for DNN training.

For DNN based acoustic modeling, we used Factored Time Delay Neural Network with additional Convolutional layers (CNN-TDNNf). 40 dimensional cepstral truncation, 3 dimensional pitch features and 100-dimensional iVectors for speaker and environment adaptation (Ghahremani et al., 2014; Miao et al., 2015) are given as input of the neural network. The CNN-TDNNf network has 15 hidden layers and a rank reduction layer . The first 6 layers are CNNs and the following 9 layers are TDNNf. The first TDNNf layer followed just after CNN layer has 256 bottleneck units and the following 8 TDNN layers consist of 1024 nodes and 128 bottleneck units.

For the monolingual experiments, we tried to tune the parameters in Kaldi's WSJ recipe using development set of Uyghur, when the Uyghur training data is less than 6 hours. However, the parameters in WSJ recipe gave us the best results even on small amount (2 hours) of Uyghur data. Thus, in all experiments, the neural network are trained with the same hyperparameters, i.e., initial learning rate (0.005), final learning rate (0.00005), minibatch-size $(128, 64)$ and training epoch (7).

For our multilingual training, we conducted 5 experiments with different sizes of Uyghur data. In each experiment, we combined the randomly selected subsets of Uyghur training data with the Turkish data and used it as training data for multilingual training. The lexicon and language models of Uyghur and Turkish are also combined. Unlike other multilingual DNN acoustic models that share only hidden layers and have language specific softmax layers, our DNN based multilingual systems have shared hidden layer as well as softmax layer. For decoding, we used language specific dictionary and language model.

3.2. Language Model

A trigram language model for Uyghur ASR was built using the training text and online broadcast text. To keep the proper size of language model and perplexity, the vocabulary of the trigram was selected so that all the words in the training data and only the words, which appeared more than ten times in the online broadcast text, were included in the language model vocabulary. The trigram contains 49k unigrams, 1.6 million bigrams and 2 million trigrams. The perplexity of the language model on evaluation set is 260.

The language model of Turkish was also prepared in similar fashion and available in our GlobalPhone corpus. The Turkish language model contained 33k unigrams, 1.6 million bigrams and 3.8 million trigrams. The perplexity of the language model on evaluation is 55.

4. Results and Discussions

4.1. Monolingual Systems

The HMM/GMM and HMM/DNN models in monolingual experiments, i.e., trained only with Uyghur data, are evaluated with evaluation set of Uyghur and the WERs are shown in Figure 2. As the size of the training data increases, the WER reduces. In each set, DNN model outperforms the best GMM model (fmllr model) and gained relative improvement ranging from 6.43% to 12.72%. The more the training data, the more relative improvement of DNN model over GMM model is achieved. With 2 hours of data, the relative improvement of DNN-model over GMM model is 6.43% and with 10 hours of data, the relative improvement is 12.72%. By increasing 2 hours of speech data in the training, on average, we gained 3.78% of relative improvement on hybrid HMM/DNN model. For example, in HMM/DNN model with 4 hours of Uyghur data, we gained 7.28% relative improvement over the HMM/DNN model with 2 hours of Uyghur data.

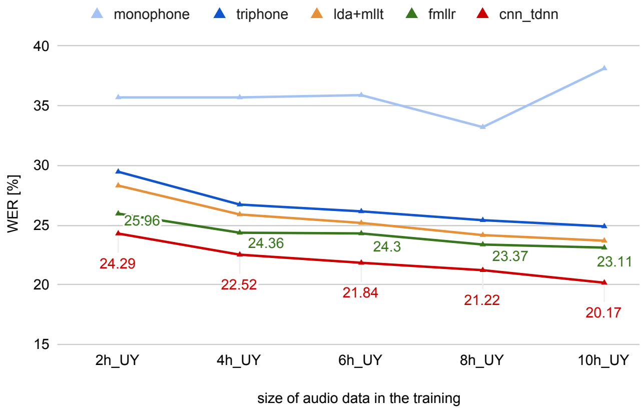

Figure 2: WER of monolingual systems trained with subsets of Uyghur data

4.2. Multilingual Systems

After training the models with combined resources, we decoded speech data (evaluation set) using the multilingual acoustic model, Uyghur pronunciation dictionary and language model. The WERs of monolingual and multilingual systems are shown in Figure 3. The blue and red lines indicate the WER of best HMM/GMM and HMM/DNN models with monolingual and multilingual training, respectively. Same as in monolingual systems, DNN models in multilingual system outperformed the GMM models. By using DNN-models, we gained 17% (on average) relative improvement over the best GMM-models (fmllr). However, the relative improvement is not so sensitive to the amount of Uyghur data used in the training. For instance, with 2 hours of Uyghur data in the training, the relative improvement of DNN-model over the GMM-model (fmllr) is 16.48% and with 10 hours of Uyghur data, it is 18.48%.

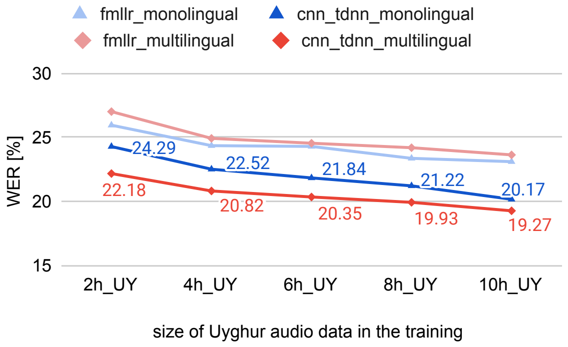

Figure 3: WER of ASR with monolingual and multilingual training

By HMM/GMM models, systems with monolingual training showed better performance than that with multilingual training in every training set. However, by HMM/DNN models, all multilingual systems outperformed monolingual ones. We compared the results of multilingual and monolingual system trained with the same amount of Uyghur data. In HMM/DNN model, we gained 2.11% of absolute WER reduction over monolingual training with 2 hours of Uyghur data, which corresponds to 8.68% relative improvement. As the size of Uyghur data in training set increases, the absolute improvement and relative improvement of speech recognition systems with multilingual training over monolingual training decreased.

With these experiments, the best ASR system for Uyghur is obtained, when 10 hours of Uyghur data is used in multilingual training. The WER of this systems is 19.27%.

We also decoded Turkish evaluation set with the multilingual acoustic model, Turkish pronunciation dictionary and language model. In Figure 4, we showed the WER of Turkish speech recognition along with the Uyghur speech recognition. It can be noted that the ASR for Turkish also benefited from multilingual training with Uyghur data. Even 2 hours of Uyghur training data lead to 0.26% relative improvement. As the amount of Uyghur data reached 8 hours, Turkish speech recognition system gained 3.85% relative improvement over the monolingual speech recognition system for Turkish (trained only with Turkish data). With the increasing size of Uyghur data, the Turkish speech recognition system is not negatively affected.

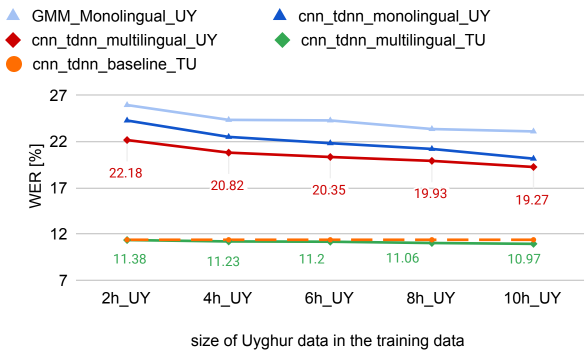

Figure 4: WER of ASR systems for Uyghur and Turkish. The dashed orange line refers to the Turkish HMM/DNN baseline system, which developed only with Turkish data and WER is 11.41%.

In all experiments, multilingual systems outperformed the monolingual systems. As shown in Figure 4, it is noticeable that absolute improvement of multilingual system over monolingual system decrease as the size of Uyghur data in the training increases. In the experiment with 10 hours of Uyghur data, the absolute improvement between multilingual and monolingual system (blue and red line) is 0.9%. From the trend of monolingual and multilingual performances (blue and red line), we can observe that the absolute improvement between multilingual and monolingual system is getting smaller, but multilingual system still outperforms monolingual system. However, we can not conclude if WER of multilingual system gets higher than that from monolingual systems, when the amount of Uyghur data becomes larger than that of Turkish data.

To explore if monolingual system outperforms multilingual system when Uyghur data becomes larger than Turkish data in the training, two additional sets of multilingual training experiments are conducted. In the first experiment, 6 hours of Turkish data is randomly selected from the Turkish training data with the constraint that all speakers are included. For the second experiment, we combined all the Turkish data (training, development and evaluation set, 17 hours 3 minutes in total). With these different sizes of Turkish data, we repeated the multilingual training experiments. The subsets of Uyghur data in these experiments remain the same as our previous multilingual training with 13 hours of Turkish data. As shown in Figure 5, the blue line presents the results from monolingual training and red line is the results from multilingual training with 13 hours of Turkish data identical as in Figure 4. The purple and green lines illustrate the results from multilingual training with 6 and 17 hours of Turkish data. In the case of larger Uyghur data used in the multilingual training than Turkish data (e.g., multilingual training with 6 hours of Turkish data and more than 6 hours of Uyghur data), multilingual systems still outperformed the monolingual system. Moreover, the multilingual system with 17 hours of Turkish data showed the best results over the other multilingual systems. From our experiments, we may conclude that Uyghur speech recognition system benefits even larger amount of Turkish data in the multilingual system.

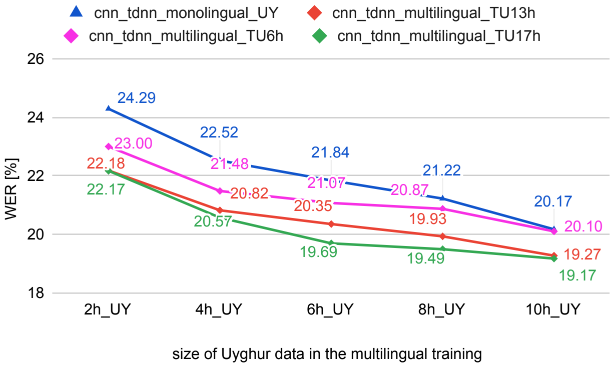

Figure 5: WER of multilingual systems trained with different sizes of Uyghur data and Turkish data.

5. Conclusion

In this paper, we developed ASR system for the low-resource language, Uyghur, by using different sizes of target language (Uyghur) data in multilingual training. We generated subsets of Uyghur data to simulate low-resource conditions and developed speech recognition systems both with monolingual and multilingual training. The results indicate that the HMM/DNN based speech recognition systems with multilingual training outperformed systems with monolingual training, despite the amount of Uyghur data used for training. In particular, we can conclude that the multilingual system for Uyghur will excel the monolingual speech recognition system, independent of size of Turkish data used in multilingual training. Both Uyghur and Turkish data benefited from each other in the multilingual training. The best speech recognition system for Uyghur is achieved by multilingual training using all Uyghur data (10 hours) and 17 hours of Turkish data and the WER is 19.17%.

6. Reference

Carki, K., Geutner, P., and Schultz, T. (2000). Turkish lvcsr: towards better speech recognition for agglutinative languages. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), volume 3, pages 1563–1566. IEEE.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., et al. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4774–4778. IEEE.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2494–2498. IEEE.

Ghahremani, P., Manohar, V., Hadian, H., Povey, D., and Khudanpur, S. (2017). Investigation of transfer learning for asr using lf-mmi trained neural networks. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 279–286. IEEE.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376. ACM.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., and Dean, J. (2013). Multilingual acoustic models using distributed deep neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8619–8623. IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 7304–7308. IEEE.

Lin, H., Deng, L., Yu, D., Gong, Y.-f., Acero, A., and Lee, C.-H. (2009). A study on multilingual acoustic modeling for large vocabulary asr. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4333–4336. IEEE.

Miao, Y., Zhang, H., and Metze, F. (2015). Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(11):1938–1949.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Sahraeian, R. and Van Compernolle, D. (2016). Using weighted model averaging in distributed multilingual dnns to improve low resource asr. *Procedia Computer Science*, 81:152–158.

Schultz, T. and Waibel, A. (2001). Language independent and language adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8126–8130. IEEE.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In Seventh International Conference on Spoken Language Processing.

Tachbelie, M. Y., Abate, S. T., and Schultz, T. (2020). Analysis of globalphone and ethiopian languages speech corpora for multilingual asr. In LREC 2020.

Vu, N. T. and Schultz, T. (2013). Multilingual multilayer perceptron for rapid language adaptation between and across language families. In Interspeech, pages 515–519.