Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 265–270
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

DNN-Based Multilingual Automatic Speech Recognition for Wolaytta using Oromo Speech

Martha Yifiru Tachbelie[1,2], Solomon Teferra Abate[1,2], Tanja Schultz[1]
[1]Cognitive Systems Lab, University of Bremen, Germany
[2]School of Information Science, Addis Ababa University, Ethiopia
abate, marthayifiru, tanja.schultz@uni-bremen.de

Abstract

It is known that Automatic Speech Recognition (ASR) is very useful for human-computer interaction in all the human languages. However, due to its requirement for a big speech corpus, which is very expensive, it has not been developed for most of the languages. Multilingual ASR (MLASR) has been suggested to share existing speech corpora among related languages to develop an ASR for languages which do not have the required speech corpora. Literature shows that phonetic relatedness goes across language families. We have, therefore, conducted experiments on MLASR taking two language families: one as source (Oromo from Cushitic) and the other as target (Wolaytta from Omotic). Using Oromo Deep Neural Network (DNN) based acoustic model, Wolaytta pronunciation dictionary and language model we have achieved Word Error Rate (WER) of 48.34% for Wolaytta. Moreover, our experiments show that adding only 30 minutes of speech data from the target language (Wolaytta) to the whole training data (22.8 hours) of the source language (Oromo) results in a relative WER reduction of 32.77%. Our results show the possibility of developing ASR system for a language, if we have pronunciation dictionary and language model, using an existing speech corpus of another language irrespective of their language family.

Keywords: Multilingual Speech Recognition, Under-resourced language, Oromo, Wolaytta

1. Introduction

Automatic Speech Recognition (ASR) is the automatic recognition and transcription of spoken language into text that can be used as text input for other systems such as information retrieval systems. Since speech is difficult to process directly in the human machine interaction, ASR technologies are important for all the human languages. As a result, a lot of research and development efforts have been exerted and lots of Automatic Speech Recognition Systems (ASRSs) have already been developed in a number of human languages. However, only insignificant number of the 7000 languages are considered.

The main reason for the limited coverage of the human languages in the development of ASRSs is that to develop an ASRS for a new language and improve the performance of the existing ones depend on the availability of speech corpus in that particular language. We do not have such corpora for a significant number of human languages, which are known to be under-resourced languages (Besacier et al., 2014). Almost all Ethiopian languages, such as Wolaytta, are under-resourced and belong to the language groups that are not benefiting from the development of spoken language technologies. To the best of our knowledge, there are only three works (Abate et al., 2020a; Tachbelie et al., 2020b; Abate et al., 2020b) towards the development of an ASRS for Oromo and Wolaytta that use at least a medium-sized speech corpora.

Multilingual Automatic Speech Recognition (MLASR) has been suggested and lots of research is being conducted in this line to solve the problem of speech corpora for under-resourced languages. MLASR system is described as a system that is able to recognize multiple languages which are presented during training(Schultz and Waibel, 2001). (Vu et al., 2014) described MLASR as a system in which at least one of the components (feature extraction, acoustic model, pronunciation dictionary, or language model) is developed using data from many different languages.

MLASR systems are particularly interesting for under-resourced languages where training data are sparse or not available at all (Schultz and Waibel, 2001). Consequently, various researches in the area of MLASR (Weng et al., 1997; Schultz and Waibel, 1998; Schultz, 2002; Kanthak and Ney, 2003; Vu et al., 2014; Müller and Waibel, 2015; Chuangsuwanich, 2016) have been conducted and a lot others are being conducted for several language groups. Especially the development of artificial neural networks (ANNs) helped to achieve better performance in the development of MLASRSs (Heigold et al., 2013; Li et al., 2019).

In our previous work (Tachbelie et al., 2020a), in which we have analyzed the similarities among GlobalPhone (Schultz et al., 2013) and Ethiopian languages (Amharic and Tigrigna from Semitic, Oromo from Cushitic and Wolaytta from Omotic), we have learned that there is high phonetic overlap among Ethiopian languages. The fact that these languages have shared phonological features is indicated in (Gutman and Avanzati, 2013) as well. From our analysis, we have learned that similarity among languages measured using their phonetic overlap crosses the boundaries of language families. Specifically, we have observed that although Oromo and Wolaytta are from different language families, there exists higher phone overlap between them than the other languages (Amharic and Tigrigna). This may be due to their geographical proximity. (Crass and Meyer, 2009) also indicated that Ethiopian languages, regardless of their language families, display areal patterns by sharing a number of similarities. Our analysis showed that 97.3% of Wolaytta phones are covered by the Oromo language while 92.3% of Oromo phones are covered by Wolaytta. Although both languages are under-resourced, Oromo is in a relatively better position than

Wolaytta. There are also a lot of other Ethiopian languages (more than 70) that are in similar or worse condition than Wolaytta with respect to language and speech resources. We wanted, therefore, to investigate the use of existing language resources to develop ASR for other Ethiopian languages. As a proof of concept, we investigated the development of Wolaytta (target language) ASR using Oromo (source language) training speech.

In this work, we present the results of different experiments we have conducted to explore the benefit we gain from MLASR approach for two languages from two different language families. First, we have conducted a cross-lingual ASR experiment where we decoded Wolaytta test speech using Oromo acoustic model (which is developed using Oromo training speech), Wolaytta language and lexical models. Second, we have developed Wolaytta ASR systems using various sizes of Wolaytta training speech (ranging from 30 minutes to 29 hours) with and without the whole amount of Oromo training speech (22.8). We have also conducted experiments to see if the source language (Oromo) can benefit from sharing training speech of the target language (Wolaytta) to improve the performance of the ASRSs.

In the following section 1.1., we give a brief description on the application of deep neural networks for the development of ASRSs. In section 2., we describe the languages considered in this paper. The speech corpora we used for the research are described in section 3. The development of the monolingual ASR using different sizes of Wolaytta training speech, which are our baseline systems, and the results achieved by the use of MLASR approach for Wolaytta using Oromo training speech are presented in section 4. Finally in section 5., we give conclusions and forward future directions.

1.1. Deep Neural Networks in ASR

Over the last 10 years, DNNs methods for ASR were developed and outperform the traditional Gaussian Mixture Model (HMM-GMM). The major factors for their superior performance are the availability of GPUs and the introduction of different types of neural network architectures such as Convolutional Neural networks (CNN) and more recently Time Delay Neural Networks (TDNN) and Factored TDNN (TDNNf).

Since 2009, DNNs are widely used in automatic speech recognition and they presented dramatic improvement in performance. Numerous studies showed hybrid HMM-DNN systems outperform the dominant HMM-GMM on the same data (Hinton et al., 2012). Currently, TDNNs, also called one-dimensional Convolutional Neural Networks, are an efficient and well-performing neural network architectures for ASR (Peddinti et al., 2015). TDNN has the ability to learn long term temporal contexts. Moreover, by using singular value decomposition (SVD) the number of parameters in TDNN models is reduced which makes them inexpensive compared to RNNs. The factored form of TDNNs (TDNNf)(Povey et al., 2018) has similar structure with TDNN, but is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal. TDNNf gives substantial improvement over TDNN and has been shown to be effective in under-resourced scenarios. We have used these state-of-the-art neural network architecture in the development of DNN based ASR systems for the Ethiopian languages.

2. Oromo and Wolaytta

More than 80 languages are spoken in Ethiopia. Ethiopian languages are divided into four major language families: Semitic, Cushitic, Omotic and Nilo-Saharan. The Semitic language family is one of the most widespread language families (with more than 20 languages) in the country. Of which, Amharic (spoken by 29.3% of the total Ethiopian population) and Tigrigna (spoken by 5.9% of the total Ethiopian population) are the most spoken languages. The Cushitic language family has also a long list of (about 22) languages spoken in Ethiopia. Amongst them, Oromo is the most widely spoken language in the country (spoken by 33.8% of the total Ethiopian population). The Omotic family has a large number of (more than 30) languages spoken in Ethiopia, one of which is Wolaytta (spoken by 2.2% of the total Ethiopian population) (CSAE, 2010).

The Cushitic and Omotic language families use Latin script for writing. In both the languages the current writers differentiate the gemminated and the non-gemminated consonants. Similarly, long and short vowels are indicated in their writing system.

Having a newly developed speech corpora (Abate et al., 2020a) for Oromo (a Cushitic language) and Wolaytta (an Omotic language), we have selected these languages to explore the application of MLASR development approach in the ANN framework.

2.1. Phonology

Although they belong to different language families, Oromo and Wolaytta share several phonetic properties including the use of long and short vowels. These languages have five similar vowels and each of the vowels in both languages has long and short variants. Having their own inventory of consonants, Oromo and Wolaytta share a number of them (see Table 1). Of course, each of the languages has its own consonants. For instance, phones ɲ and x are used in Oromo but not in Wolaytta while phone ʒ is used in Wolaytta but not in Oromo.

Almost all the consonants of these languages occur in both single and gemminated forms. The other common phonetic feature of these languages is the use of tones which makes both of them tonal languages. However, in this study we did not differentiate between vowels of different tones since the writing system does not show the tones of the vowels and the pronunciation dictionaries for our study have been generated automatically from the text.

Language	Consonants (IPA)	Vowels (IPA)
Oromo	b d ɗ f g h j k k’ l m n ɲ p p’ r s ʃ t t’ tʃ tʃ’ ʤ v w x z ʔ	a e i o u aː eː iː oː uː
Wolaytta	b d ɗ f g h j k k’ l m n p p’ r s ʃ t t’ tʃ tʃ’ ʤ w z ʒ ʔ	a e i o u aː eː iː oː uː

Table 1: Oromo and Wolaytta phones

2.2. Morphology

Reflecting the morphological nature of their language families, Oromo and Wolaytta are not as simple as English and not as complex as the Semitic language families. In both Oromo and Wolaytta nominals are inflected for number, gender, case and definiteness and verbs are inflected for person, number, gender, tense, aspect and mood (Griefenow-Mewis, 2001). Unlike the Semitic languages, which allow prefixing, Oromo and Wolaytta are suffixing languages. In these languages words can be generated from stems recursively by adding suffixes only.

3. The Speech Corpora

It is known that the Ethiopian languages, specially Oromo and Wolaytta are under-resourced. As a result, all of the previous works conducted towards the development of ASRSs for these languages are based on limited amounts of speech data. It is only recently that a work on the development of four standard medium-sized read speech corpora (Abate et al., 2020a) has been conducted for four Ethiopian languages including Oromo and Wolaytta. For a country like Ethiopia with more than 80 languages, unless a technological solution is used, it looks hopeless to have equivalent speech corpora for all its languages.

In this work, we have used the existing speech corpora of Oromo (Abate et al., 2020a) to find out a solution for the development of an ASRS for an under-resourced language, Wolaytta. We considered Oromo as a source and Wolaytta as a target language considering the fact that there are more previous works conducted for Oromo, such as (Gelana, 2016; Gutu, 2016) than what we have for Wolaytta. We hope that our findings will be extended to solve the problems of the other Ethiopian languages that fall under four different language families.

4. Multilingual ASR for Wolaytta

4.1. Development of ASR Systems for Wolaytta

Although the aim of our current work is to explore the development of MLASR for Wolaytta as a target language using Oromo training speech (as a source language), we have developed different monolingual GMM- and DNN-based ASRSs for Wolaytta using different sizes of Wolaytta speech corpus for comparison purposes. The description of the procedures we followed is presented in sub-section 4.1.1..

4.1.1. Acoustic, Lexical and Language Models

To build reference AMs that use different sizes of training speech, we have splitted the Wolaytta training speech into 11 clusters: with 30 minutes, 1, 2, 4, 6, 8, 10, 15, 20, 25 and 29 (all) hours of speech length. We have selected roughly equal number of utterances from each speaker randomly for each of these clusters. Each of them has been used to train different AMs.

All the AMs have been built in a similar fashion using Kaldi ASR toolkit (Povey et al., 2011). We have built context dependent HMM-GMM based AM using 39 dimensional mel-frequency cepstral coefficients (MFCCs) to each of which cepstral mean and variance normalization (CMVN) is applied. The AM uses a fully-continuous 3-state left-to-right HMM. Then we did Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation for each of the models. Then Speaker Adaptive Training (SAT) has been done using an offline transform, feature space Maximum Likelihood Linear Regression (fMLLR). We did tuning to find the best number of states and Gaussians for different sizes of the training data.

To train the DNN-based AMs, we have used the best HMM-GMM models to get alignments and the same training speech used to train HMM-GMM models. But we have applied a three-fold data augmentation (Ko et al., 2015) prior to the extraction of 40-dimensional MFCCs without derivatives, 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation. The neural network architecture we used is Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf) according to the standard Kaldi WSJ recipe. The Neural network has 15 hidden layers (6 CNN followed by 9 TDNNf) and a rank reduction layer. The number of units in the TDDNf consists of 1024 and 128 bottleneck units except for the TDNNf layer immediately following the CNN layers which has 256 bottleneck units.

The list of word entries both for training and decoding lexicons have been extracted from the training speech transcription in both the source and target languages. Using the nature of writing system that indicates gemminated and non-gemminated consonants as well as the long and short vowels, we have generated the pronunciation of these words automatically. However, since the tones are not indicated in written form of both languages, we did not consider tones in the current pronunciation dictionaries.

For the development of the LMs we have used the text used in (Abate et al., 2020a). We have developed trigram LMs using the SRILM toolkit (Stolcke, 2002). The LMs are smoothed with unmodified Kneser-Ney smoothing techniques (Chen and Goodman, 1996) and made open by including a special unknown word token. LM probabilities are computed for the lexicon of the training transcription.

4.1.2. Evaluation Results

We have evaluated all AMs trained with different sizes of training speech using the same test set (1:45 hours of speech recorded from four speakers who read a total of 578 utterances), pronunciation dictionary and language model. The performance of the systems is given in Figure 1. These results are our reference points or baselines for the results achieved by using only the source language, and combined with different amounts of target language's training speech.

As we can observe from Figure 1, obviously, the WER reduces with the additional training speech in almost all the AMs. The DNN-based systems outperform the HMM-GMM-based ones regardless of the size of the training speech, except for 30 minutes. The DNN-based AMs has brought a relative WER reductions that range from 9.03% (with 1 hour) to 31.45% (with all the training speech). The best system developed using all the available training speech has achieved a WER of 23.23% with the DNN-based AM.

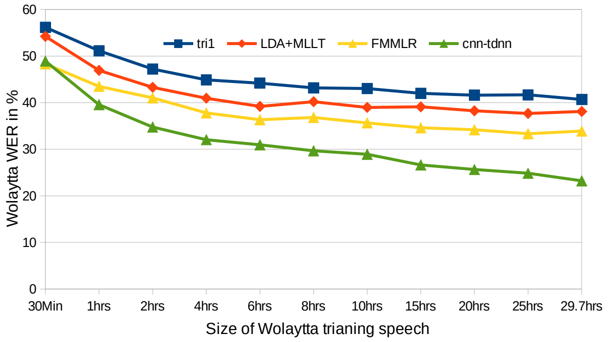

Figure 1: Wolaytta WERs with different sizes of Wolaytta training speech

4.2. Use of Oromo Speech for Wolaytta ASR

First we have decoded the Wolaytta evaluation test speech using a DNN-based Oromo AM (trained using all the training speech of the Oromo corpus), Wolaytta pronunciation dictionary and Wolaytta language model and achieved a WER of 48.34%. For this purpose we needed to map the Wolaytta phones that are not found in Oromo to the nearest possible Oromo phones (see Table 2).

Wolaytta Phones (IPA)	Mapped Oromo Phones(IPA)	Remarks Remarks
7 (ʔ)	hh (ʔ)	Same IPA
zh (ʒ)	z (z)	Different IPA
zz (z:)	z (z) z (z)	Double to single mapping
ssh (ʃ:)	sh (ʃ) sh (ʃ)	Double to single mapping
hhhh (ʔ:)	hh (ʔ) hh (ʔ)	Double to single mapping

Table 2: Wolaytta phones mapped to Oromo phones

We have, then, conducted experiments to see the benefits it gets from additional Wolaytta speech incrementally starting from 30 minutes to the whole training speech. The evaluation of all the systems is done using the same evaluation set. The results are presented in Figure 2.

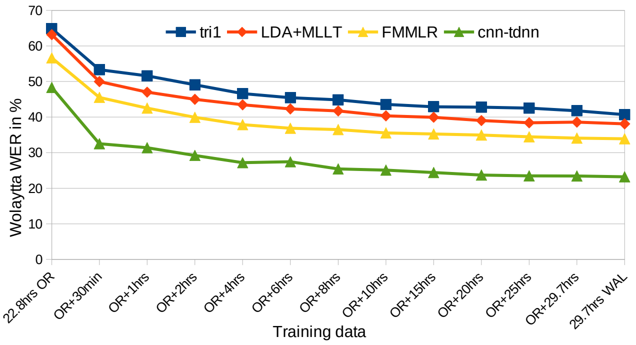

Figure 2: Wolaytta WERs with different sizes of Wolaytta training speech added to the whole training speech of Oromo

The results in Figure 2 show that performance improvement can be obtained by adding training speech from the target language. As we add more and more training speech from the target language, the improvement in performance reduces. A relative WER reduction of 32.77% has been achieved as a result of adding only 30 minutes of training speech from the target language. That means the WER we could achieve by using only the source language's training speech has been reduced from 48.34% to 32.5% by adding only 30 minutes training speech of the target language that is randomly selected from all the speakers (76) of the target language.

Our results also show that instead of using only small amount of monolingual training speech in the development of an ASRS, specially in the DNN framework, the use of speech data from other related languages bring performance improvement. We have presented this improvement in Figure 3 that shows the comparison of WERs of ASRSs developed using Wolaytta training speech only and that of the ASRSs developed using different sizes of training speech from Wolaytta combined with all (22.8 hours) Oromo training speech. As it can be seen from the Figure, by adding only 30 minutes of Wolaytta training speech to all of the Oromo training speech, we have achieved a relative WER reduction of 33.55% and 5.52% when 25 hours of Wolaytta training speech is added.

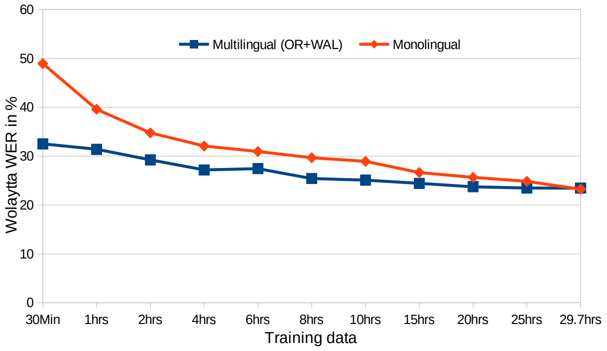

Figure 3: Wolaytta WERs with different sizes of Wolaytta with and without the Oromo speech

4.3. Evaluation of Multilingual Acoustic Models for Oromo

We have decoded Oromo test set using the acoustic models (Wolaytta only AM and MLASR AMs) discussed in the previous sections, Oromo pronunciation dictionary and Oromo language model developed by (Abate et al., 2020a). The results presented in Figure 4 show that we have achieved a WER of 49.25% using the DNN-based AM developed using 29.7 hours of Wolaytta training speech. The performance of MLASR systems on Oromo test set brought slight WER reductions compared to the best WER obtained from a system that is developed using Oromo training speech only. The relative WER reductions we have obtained range from 1.27% (gained from the addition of 10 hours of Wolaytta speech) to 3.31% (gained from the addition of 25 hours of Wolaytta speech). We could observe that adding 30 minutes to 8 hours of Wolaytta training speech has negatively affected Oromo ASR.

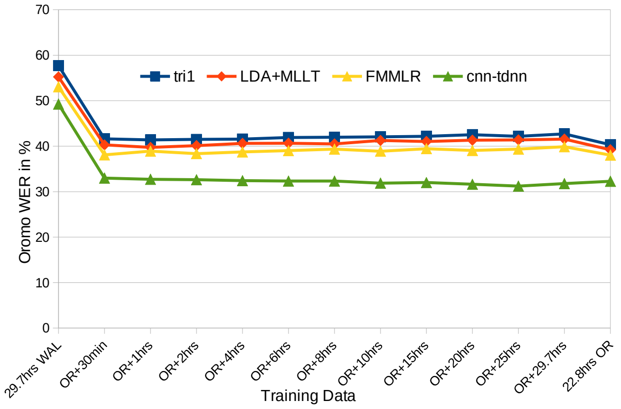

Figure 4: Oromo WERs with different sizes of Wolaytta training speech with and without the Oromo speech

5. Conclusion and Way Forward

In this paper, we have presented the experiments conducted on the development of multilingual ASRs across language families taking Oromo and Wolaytta as source and target languages, respectively. We have achieved a WER of 48.34% for Wolaytta without any training speech from it. By adding only 30 minutes of speech data from Wolaytta to the whole training data of the source language (Oromo) we have achieved a relative WER reduction of 32.77%. The ASRSs developed using all the training speech (22.8 hours) of the source language together with different sizes of training speech from the target language outperformed the ASRSs developed using training speech of the respective size from the target language only. The observed relative WER reductions range from 33.55% (achieved when training speech of Oromo plus only 30 minutes of Wolaytta is used) to 5.52% (achieved when training speech of Oromo plus 25 hours of Wolaytta is used). Based on our results, we conclude that it is possible to develop an ASRS with reasonable performance for a language using speech data of another language, irrespective of its language family, provided that we have a decoding pronunciation dictionary and a language model. We, therefore, recommend the development of a decoding pronunciation dictionary and a language model for the other Ethiopian languages so that they can benefit from the development of MLASRSs using the speech corpora of other languages.

6. Acknowledgment

We would like to express our gratitude to the Alexander von Humboldt Foundation for funding the research stay at the Cognitive Systems Lab (CSL) of the University of Bremen.

7. Bibliographical References

Abate, S. T., Tachbelie, M. Y., Melese, M., Abera, H., Abebe, T., Mulugeta, W., Assabie, Y., Meshesha, M., Atinafu, S., and Ephrem, B. (2020a). Large vocabulary read speech corpora for four ethiopian languages : Amharic, tigrigna, oromo and wolaytta. In *LREC 2020*.

Abate, S. T., Tachbelie, M. Y., and Schultz, T. (2020b). Deep neural networks based automatic speech recognition for four ethiopian languages. In *ICASSP 2020*.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

Chuangsuwanich, E. (2016). *Multilingual techniques for low resource automatic speech recognition*. Ph.D. thesis.

Crass, J. and Meyer, R. (2009). *Introduction*. Rüdiger Köppe Verlag, Köln.

CSAE. (2010). The 2007 population and housing census.

Gelana, K. (2016). *A Large Vocabulary, Speaker-Independent, Continuous Speech Recognition System for Afaan Oromo: Using Broadcast News Speech Corpus*. Ph.D. thesis, Addis Ababa University.

Griefenow-Mewis, C. (2001). *A Grammatical Sketch of Written Oromo*.

Gutman, A. and Avanzati, B. (2013). Languages of ethiopia and eritrea.

Gutu, Y. G. (2016). *A Continuous, Speaker Independent Speech Recognizer for Afaan Oroomoo: Afaan Oroomoo Speech Recognition Using HMM Model*. Ph.D. thesis, Addis Ababa University.

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.

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.

Kanthak, S. and Ney, H. (2003). Multilingual acoustic modeling using graphemes. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1145–1148.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *INTERSPEECH*.

Li, X., Dalmia, S., Black, A., and Metze, F. (2019). Multilingual speech recognition with corpus relatedness sampling, 08.

Müller, M. and Waibel, A. H. (2015). Using language adaptive deep neural networks for improved multilingual speech recognition.

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

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.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747.

Schultz, T. and Waibel, A. (1998). Multilingual and crosslingual speech recognition. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, pages 259–262.

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

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text and speech database in 20 languages. In *ICASSP*.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In John H. L. Hansen et al., editors, *INTERSPEECH*. ISCA.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, pages 901–904.

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

Tachbelie, M. Y., Abulimiti, A., Abate, S. T., and Schultz, T. (2020b). Dnn-based speech recognition for globalphone languages. In *ICASSP 2020*.

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

Weng, F., Bratt, H., Neumeyer, L., and Stolcke, A. (1997). A study of multilingual speech recognition. In *EUROSPEECH*.