Multilingual Acoustic and Language Modeling for Ethio-Semitic Languages

Solomon Teferra Abate[1,2], *Martha Yifiru Tachbelie*[1,2], *Tanja Schultz*[2]

[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

Development of Multilingual Automatic Speech Recognition (ASR) systems enables to share existing speech and text corpora among languages. We have conducted experiments on the development of multilingual Acoustic Models (AM) and Language Models (LM) for Tigrigna. Using Amharic Deep Neural Network (DNN) AM, Tigrigna pronunciation dictionary and trigram LM, we achieved a Word Error Rate (WER) of 30.9% for Tigrigna. Adding training speech from the target language (Tigrigna) to the whole training speech of the donor language (Amharic) continuously reduces WER with the amount of added data. We have also developed different (including recurrent neural networks based) multilingual LMs and achieved a relative WER reduction of 3.56% compared to the use of monolingual trigram LMs. Considering scarcity of computational resources to decode with very large vocabularies, we have also experimented on the use of morphemes as pronunciation and language modeling units. We have achieved character error rate (CER) of 7.9% which is relatively lower by 38.3% to 1.3% than the CER of the word-based models of smaller vocabularies than 162k. Our results show the possibility of developing ASR system for an Ethio-Semitic language using an existing speech and text corpora of another language in the family.

Index Terms: Multilingual Acoustic model, Multilingual Language model, Less-resourced language, Amharic, Tigrigna

1. Introduction

The development of an Automatic Speech Recognition (ASR) system requires developing pronunciation, acoustic and language models. The development of an acoustic model (AM) requires a speech corpus while there should be a large collection of text corpus for the development of a language model (LM). These resources have been developed for only a few of the 7000 languages in the world. Languages are categorized as well-resourced and under-resourced based on availability of language resources in a language [1]. Almost all Ethiopian languages are under-resourced and belong to the language groups that are not benefiting from the recent development of spoken language technologies.

The Ethiopian languages are not only under-resourced but are also morphologically complex, which makes the development of ASR more difficult and demand even more language resources. Development of language resources for all languages is costly. Researchers have been, therefore, looking for solutions to develop ASRs for under-resourced languages through Multilingual Automatic Speech Recognition (MLASR). MLASR system is a system that is able to recognize multiple languages which are presented during training [2]. As described in [3] MLASR is a system in which at least one of the components (feature extraction, AM, pronunciation dictionary, or LM) is developed using data from many different languages. Literature shows that the challenge of morphological complexity and the consequent problem of high out-of-vocabulary (OOV) rate has been tackled in three ways: use of large decoding vocabulary [4], multilingual (ML) language modeling [5, 6] and use of sub-word units for pronunciation and language modeling [7, 8]

Although MLASR systems are useful in a number of ways, including the development of language agnostic speech technologies [9], they are particularly interesting for under-resourced languages where training data for the development of any one or all of its components (acoustic, pronunciation and/or language models) are sparse or not available at all [2]. Consequently, various studies in MLASR [10, 11, 12, 13, 3, 14, 15, 16, 4] have been conducted for several language groups. The research trend, as it is the case in all other machine learning research, shows that the use of artificial neural networks (ANNs) results in better performance in the development of MLASR systems [17, 18, 19].

In this work, we present the results of different experiments conducted towards the development of ML acoustic and language models for Ethio-Semitic languages as a proof of concept. In this language family, there are lots of morphologically complex languages for which we do not have speech corpora for acoustic modeling and text corpora for language modeling. To use ANNs for acoustic and language modeling, we do not have large training speech and text data for any one of the Ethio-Semitic languages. We have, therefore, conducted experiments to see if we can develop and also improve the performance of the ASR system of a target language by sharing training speech for acoustic modeling and text for language modeling. Sharing training text among languages for the development of LMs is not as common as the idea of sharing speech for the development of AMs. But recently literature [6] showed that it is a promising direction towards developing strong LMs for under-resourced languages.

Since morphological complexity of this language family is one of the challenges in language modeling, we have also conducted experiments on the use of sub-word pronunciation and language modeling units by segmenting words into sub-word units using Morfessor [20].

In section 2, we give a brief description on the application of deep neural networks (DNN) for acoustic and language modeling. In section 3, we describe morphological and phonological relationships between the languages considered. The speech and text corpora we used for the research are described in section 4. The baseline AMs, our experiments on the development of ML AMs and language modeling experiments along with the coresponding results are presented in sections 5 and 6, respectively. We conclude in section 7.

2. DNN for Acoustic and Language Model

Since 2009, DNNs are widely used for the development of AMs and their usage results in significant improvement in performance. Numerous studies showed hybrid HMM-DNN systems

outperform the dominant HMM-GMM on the same data [21]. Currently, Time delay Neural Networks (TDNNs), also called one-dimensional Convolutional Neural Networks, are efficient and well-performing neural network architectures for ASR [22]. 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)[23] 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 AMs for the Ethio-Semitic languages.

Lately, Recurrent Neural Network Language Models (RNNLMs) introduced by [19] in 2010 replaced the standard n-gram techniques. Since then the use of RNNLM has brought WER reduction for ASR systems. RNNLMs are mostly used in rescoring a precompiled decoding graph that is generated in the first pass using n-gram LMs. The run time and performance of RNNLMs have been further improved by different researchers such as [24]. Consequently, they are now becoming more popular component of ASR system.

3. Ethiopian Languages

In this study, we considered the Ethio-Semitic language family, which consists of a number of languages namely, Amharic, Tigrigna, Geez, Tigre, Argobba, Harari, Gafat, Soddo and other Gurage languages. For our research we have taken Amharic and Tigrigna. Both have a considerable number of native speakers. Amharic is spoken by more than 27 million people which makes it the second most spoken Semitic language while Tigrigna is spoken by 9 million people. These languages have different functions in Ethiopia. Amharic is the working language of the Federal Government and the Amhara regional state. Tigrigna is the working language of Tigray regional state. It is also the language of Eritrea. Apart from this, they serve as medium of instructions in primary and secondary schools. A number of governmental websites are available in Amharic and Tigrigna. These languages are available in electronic media like news, blogs and social media. Currently, Google offers a searching capability in Amharic and Tigrigna. Due to the larger number of speakers, wider geographical coverage and more functions in Ethiopia, these two languages have significant impact on the other Ethiopian languages. We, therefore, believe that the concept proofed for these two languages can be applied for the other Ethiopian languages.

3.1. Writing System

The writing system of the Ethio-Semitic languages is Ethiopic which is syllabic where each character represents a consonant plus a vowel. The feature of the writing system is that each character gets its basic shape from the consonant of the syllable, and the vowel is represented through a systematic modifications of this shape. The script is used to write a number of other Ethiopian languages. This writing system does not show consonant gemination, the presence and absence of the epenthetic vowel and the glottal stop consonant.

3.2. Morphology

Reflecting their Semitic language morphology, Amharic [25] and Tigrigna [26], make use of the root and pattern system. In these languages, a root (called radical) is a set of consonants which bears the basic meaning of the lexical item whereas a pattern is composed of a set of vowels inserted between the consonants of the root. These vowel patters together with affixes, result in derived words. Such derivational process makes these languages morphologically complex. Furthermore, an orthographic word attaches some syntactic words like prepositions, conjunctions, negation, etc. which results in a large amount of word forms. In these languages, nominals are inflected for number, gender, definiteness and case whereas verbs are inflected for person, number, gender, tense, aspect, and mood.

3.3. Phonology

Amharic and Tigrigna share a lot of phones. All the 35 phonemes (28 consonants and 7 vowels) used in Amharic are found in Tigrigna that has four more phonemes. In both languages there are labialized phones arguably represented either as a set of labialized consonants or a set of labialized vowels. In this work, we have represented them as labialized vowels:uə, ui, ua, ue, uɨ. The four Tigrigna sounds that are not found in Amharic are ʕ, ħ, x and x́. The glottalized or ejective tˈ kˈ pˈ tsˈ tʃˈ sounds are found in both Amharic and Tigrigna [25]. Long consonants or geminated consonants, are clearly pronounced and bring semantic difference in these languages. Both languages have seven vowels: ə, u, i, a, e, ɨ, o.

4. The Speech Corpora

To the best of our knowledge, there is only one known standard medium-sized read speech corpus [27] for Amharic and one similar speech corpus for Tigrigna [28]. The Tigrigna speech corpus is not yet made available to the research community.

Recently, four medium-sized read speech corpora have been developed for four Ethiopian languages including Amharic and Tigrigna [29]. Given the time and cost consuming process of developing such corpora, it is unlikely that such corpora will be developed for all 80 Ethiopian languages any time soon.

In this work, we have used the existing speech corpora of Amharic [27] and Tigrigna [29] to investigate the concept of using exiting resources from a language family for the development of ASR for another language in the same family. We considered the well researched language of Amharic as a donor or source and Tigrigna as a target language. We hope that this proof-of-concept study will provide a helpful recipe for the rapid development of speech and language technologies in the many Ethiopian languages.

The Amharic corpus consists of development and evaluation test sets of 760 utterances with about 1.5 hours of speech, each. The Tigrigna development and evaluation test sets, consisting of 511 and 507 utterances, respectively, a total speech of 1 hour, each are held out from the total recording. We have considered gender balance in selecting these test sets.

5. Multilingual Acoustic Modeling

5.1. The Baseline Pronunciation and Acoustic Models

We have considered the AMs presented in our previous work [4] as baselines. In [4] we have used different sizes of Tigrigna decoding vocabularies ranging from 32.5k to 299k. The word entries for the development of the decoding vocabularies have been extracted (based on their frequencies) from all the available texts (about 4 Million word tokens for each language) that are prepared for language modeling, including the training transcriptions. Using the syllabic nature of the writing system, we generated pronunciations of these words automatically.

All the AMs have been built in the same way using the Kaldi ASR toolkit [30]. We have built context dependent HMM-GMM based AM using 39 dimensional mel-frequency

cepstral coeffcinents (MFCCs) to each of which cepstral mean and variance normalization (CMVN) have been applied. The AM uses a fully-continuous 3-state left-to-right HMM. We did Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation. The Speaker Adaptive Training (SAT) has been done using an offline transform, feature space Maximum Likelihood Linear Regression (fMLLR).

To train the DNN-based AMs, we have used the best HMM-GMM model to get alignments and the same training speech used to train HMM-GMM model. We have applied a three-fold data augmentation [31] before extracting 40-dimensional MFCCs without derivatives, 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation. The DNN architecture we used is Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf) according to the standard Kaldi WSJ recipe that 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 following the CNN layers which has 256 bottleneck units.

5.2. Development of Monolingual Acoustic Models for Tigrigna Using Different Amounts of Training Speech

To simulate different states of resource scarcity, we have splitted the Tigrigna training speech into samples of 1, 2, 4, 6, 8, 10, 15 and 22 (all) hours. We have selected a roughly equal number of utterances from each speaker randomly for each of these samples. Considering the size of the training data, we have experimented with different DNN architectures by reducing the number of hidden layers. But insignificant WER reduction is observed from the use of 12 hidden layers only for systems developed with 1 hour and 2 hours of training speech. Therefore, the baseline architecture is used to develop different GMM- and DNN-based AMs for Tigrigna using each of these samples.

The performance of these AMs have been evaluated using the largest Tigrigna decoding pronunciation dictionary and the baseline Tigrigna trigram LM. As presented in Figure 1, the performance of the AMs improve as the size of the training speech increases, especially the DNN-based AMs. The relative WER reduction obtained by using DDN AMs over the GMM AMs grew (from 5.43% for 1 hour of training speech) to 26.40% when 15 hours of training speech is used. This clearly shows the data greedy nature of DNN-based AMs. However, we have learnt that it is possible to develop DNN-based AM, using as small training speech as 1 hour and achieve a WER of 25.77% that goes below 20% with only 6 hours of training speech.

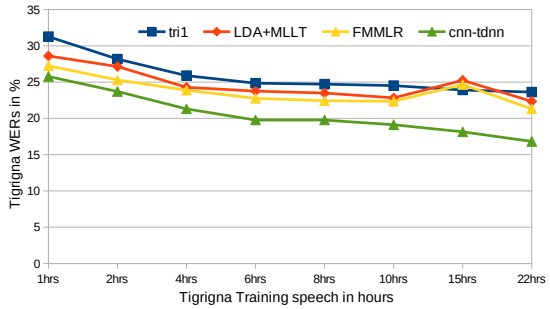

Figure 1: *Tigrigna WERs with different amounts of training speech.*

5.3. Development of Multilingual Acoustic Models for Tigrigna Using Amharic Corpus

To consider the most challenging case of zero training speech [32] for the target language, we decoded the Tigrigna evaluation set using DNN-based Amharic AM (trained using all the training speech of the Amharic corpus), Tigrigna pronunciation dictionary (with OOV rate of 4.89%) and Tigrigna trigram LM (with perplexity 172.42) and achieved a WER of 30.9%. For this purpose, we needed to map the Tigrigna phones (ʕ, ħ, x and x́) that are not found in Amharic to the phonetically nearest possible Amharic phones (ʔ, q, h, and h). From the analysis of the errors, we could observe that most of the recognition errors are related to this mapping of the four Tigrigna phones that are not available in the Amharic training speech.

To see the effect of such difference in phone sets between languages, we have decoded the Amharic evaluation set using Tigrigna DNN-based AM, Amharic 310k decoding pronunciation dictionary, and Amharic trigram LM (which has 3.06% OOV rate and perplexity of 41.2, reflecting a closed domain task) and achieved a WER of 9.68%. Again to see the performance of such a zero training speech on an open domain task, we have decoded the evaluation test set of the newly developed Amharic corpus [29] which consists of 508 utterances and 1.25 hours of speech. The decoding pronunciation dictionary of this system is 323k that has OOV rate of 6.21% and the LM perplexity is 241.26 on the new test text. Clearly showing the domain effect, the WER of our system increased to 26.01%. The lower WER we got for Amharic than the Tigrigna WER (30.9%) can, at least partially, be attributed to the full coverage of Amharic phones by the Tigrigna training speech.

Taking the performance of Tigrigna ASR system with the zero Tigrigna training speech as a baseline, we have conducted experiments to see the benefits we get from adding Tigrigna training speech incrementally starting from 1 hour to 22 hours. The ML AMs are trained using ML mix approach [33]. The results are presented in Figure 2 and show the trend in performance improvement that is obtained by adding training speech from the target language (Tigrigna).

We have observed that as we add more and more training speech of the target language, the degree of improvement in performance reduces. A relative WER reduction of 45.79% (over the WER of the zero training speech Tigrigna ASR system) has been achieved as a result of adding only 1 hour of Tigrigna training speech. The gain from the last 12 hours of additional training speech is only 6% relative WER reduction. This gives us an understanding that less than 10 hours of training speech for a new target language will enable us develop an AM that has a competitive performance of a monolingual AM developed using more than 20 hours of monolingual training speech.

Our results show that instead of using only small amount of monolingual training speech in the development of an ASR system, the use of speech data from other related languages bring performance improvement. As it can be seen from the results presented in Figure 1 for DNN-based monolingual AM and Figure 2 for DNN-based ML AM, relative WER reductions ranging from 21.27% to 0.96% has been achieved as a result of using the Amharic speech data to train Tigrigna AMs.

6. Language Modeling Experiments

We have also conducted experiments towards the development of ML LMs applying two language modeling approaches: n-gram and RNNLM. We have considered the LMs presented in our previous work [4] as baselines. These LMs are open vocabulary trigram LMs developed using the SRILM toolkit [34]

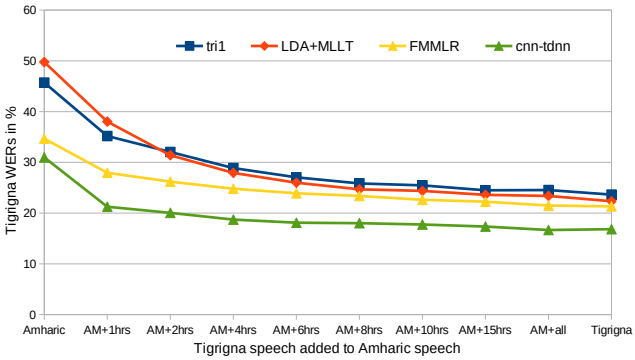

Figure 2: *Tigrigna WERs with different sizes of Tigrigna added to the whole Amharic training speech.*

with unmodified Kneser-Ney smoothing techniques [35]. The LM probabilities are computed for the words included in the decoding pronunciation dictionaries.

To have baseline LMs for RNNLM approach, we have developed monolingual RNNLMs using the kaldi toolkit. We have used a network of 5 layers, which is composed of 3 TDNN and 2 LSTM layers. The embedding dimension that works best on our data is 1000 in each of the hidden layers. We have experimented with different learning epochs and found out that learning with epoch 10 is optimal. The LMs are evaluated in a pruned lattice rescoring [24] framework on lattices generated using the baseline trigram LMs. The results of the best performing LMs with epoch 10 are presented in Table 1.

6.1. Multilingual Language Modeling

We have developed the ML LMs using the different Tigrigna vocabularies, Amharic training texts and Tigrigna training texts. We experimented with two-fold data augmentation for the target language text and different weights (higher weights assigned to the target language, for example 0.75 for the target and 0.25 for the source). However, the use of equal weight for source and target languages data without data augmentation outperformed. Only the best results are presented in Table 1. The performance

Table 1: *Performance of Mono- and Multilingual LMs*

Vocab	OOV	Mono WER		ML WER	
	rates	trigram	rnnlm	trigram	rnnlm
32k	15.41	26.94	27.02	26.81	26.62
65k	11.01	22.55	22.66	22.42	22.48
97k	8.89	20.81	20.54	20.54	20.23
130k	7.41	19.59	19.37	19.32	19.04
162k	7.01	19.11	18.85	18.91	18.43
195k	6.57	18.67	18.43	18.39	18.10
227k	6.03	17.94	17.75	17.84	17.53
260k	5.68	17.55	17.53	17.45	17.10
292k	5.00	16.92	16.74	16.85	16.51
299k	4.89	16.82	16.64	16.71	**16.35**

of Tigrigna monolingual RNNLMs showed (except for the two smallest, 32.5k and 65k, vocabularies) improvement over the performance of the baseline trigram LMs. The results of the ML LMs also show WER reductions. This shows that Tigrigna has benefited from the use of Amharic text in language modeling in both approaches (n-gram and RNNLM).

6.2. Morphem-based Language Modeling

The best performing ASR system achieved a WER of 16.35% using 299k decoding vocabulary, the biggest ML LM developed with RNNLM and two pass decoding. The development of such a system is computationally expensive, especially for developing countries where not only language resources but also computational resources are scarce. An alternative approach for languages of such countries, especially for under-resourced and morphologically complex languages, is the use of sub-word units such as morphemes in language and pronunciation modeling. Thus, we conducted experiments on the use of morphemes (obtained using Morfessor) for pronunciation and language modeling. Our results show that this modeling unit reduces the total size of our decoding vocabulary from 299k to 66.5k and the OOV rate from 4.89% to 0.02%. Using the segmented Tigrigna training text and the morph vocabulary, we have developed different n-gram LMs (n=3 to 9) and used it to decode the Tigrigna evaluation set using the monolingual DNN-based AMs. For the purpose of comparison we computed character error rates (CERs) of the morph-based hypotheses and the respective hypotheses of the baseline systems. The CER of the best (using 6-gram LM) morpheme-based ASR system is 7.9%. As can be seen from Table 2 (last row), we achieved relative CER reductions (MorLM CER Red) ranging from 1.3% to 38.3% from the use of of morpheme-based LM over the word-based LMs that use vocabularies of less than 162k.

Table 2: *Relative CER Reduction of Morph-based LM*

Vocab	32k	65k	97k	130k	162k
CER of Word LM	12.8	10.3	8.9	8	7.8
MorLM CER Red	**38.3**	23.3	11.2	1.3	-1.3

7. Conclusions

In this paper, we have presented the experiments conducted on the development of a MLASR taking Amharic and Tigrigna as cases. The results we have achieved show that it is possible to develop an ASR system for an Ethio-Semitic language like Tigrigna using an existing speech corpora of another language in this language family such as Amharic. We have achieved a WER of 30.9% without any training speech from the target language (Tigrigna). We have also observed that adding speech data from the target language to that of the source language (Amharic) continuously reduces the WER up to relative reduction of 21.27%. The ASR systems developed using speech data from a source language together with different size of speech data of the target language outperformed the ASR systems developed using the target language data only. We have also conducted experiments on the way to develop strong LMs for the target language using texts of the donor language by applying RNNLM and achieved WER reductions. Considering scarcity of computational resources to decode using very large vocabularies, we have also developed morpheme-based pronunciation and language models. Our results showed that the morpheme-based systems with only 66.5k vocabularies achieve comparable CER with very large word-based decoding vocabularies.

8. Acknowledgment

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

9. References

[1] 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.

[2] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001.

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

[4] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Deep neural networks based automatic speech recognition for four ethiopian languages," in *ICASSP 2020*, 2020.

[5] A. Ragni, E. Dakin, X. Chen, M. J. Gales, and K. M. Knill, "Multi-language neural network language models," in *Interspeech 2016*, 2016, pp. 3042–3046.

[6] A. Abulimiti and T. Schultz, "Building language models for morphological rich low-resource languagesusing data from related donor languages: the case of uyghur," in *roceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, 2020, p. 271276.

[7] Martha Yifiru Tachbelie and Solomon Teferra Abate and Wolfgang Menzel, "Morpheme-based automatic speech recognition for a morphologically rich language - amharic," in *2nd Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2010*, 2010, pp. 68–73.

[8] M. Y. Tachbelie, S. T. Abate, and L. Besacier, "Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic," *Speech Commun.*, vol. 56, pp. 181–194, jan 2014.

[9] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," in *ICASSP 2020*, 2020, pp. 8239–8243.

[10] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *EUROSPEECH*, 1997.

[11] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[12] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *INTERSPEECH*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[13] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," in *IN PROCEEDINGS OF EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY*, 2003, pp. 1145–1148.

[14] M. Müller and A. H. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," 2015.

[15] E. Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, 2016.

[16] N. Beringer and F. Schiel, "The quality of multilingual automatic segmentation using german maus," in *INTERSPEECH*, 2000.

[17] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8619–8623.

[18] X. Li, S. Dalmia, A. Black, and F. Metze, "Multilingual speech recognition with corpus relatedness sampling," 08 2019.

[19] T. Mikolov, M. Karafit, L. Burget, J. Cernock, and S. Khudanpur, "Recurrent neural network based language model," vol. 2, 01 2010, pp. 1045–1048.

[20] O. Kohonen, S. Virpioja, and K. Lagus, "Semi supervised learning of concatenative morphology," in *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Association for Computational Linguistics, July 2010, pp. 78–86.

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

[22] 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.

[23] 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 *Interspeech*, 2018, pp. 3743–3747.

[24] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," *ICASSP*, pp. 5929–5933, 2018.

[25] W. Leslau, *Introductory grammar of Amharic*, ser. Porta linguarum orientalium, neue Serie, Bd. 21. Wiesbaden: Harrassowitz, 2000.

[26] T. Y. Tesfay, *A modern grammar of Tigrinya*. Rome: Tipografia U. Detti, 2002.

[27] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *INTERSPEECH 2005*, 2005, pp. 1601–1604.

[28] H. Abera and S. Hailemariam, "Design of a Tigrinya language speech corpus for speech recognition," in *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 78–82.

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

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

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[32] M. Prasad, D. van Esch, S. Ritchie, and J. F. Mortensen, "Building large-vocabulary asr systems for languages without any audio training data," in *Proceedings of Interspeech 2019*, 2019.

[33] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Elsevier Academic Press, 2006.

[34] A. Stolcke, "Srilm – an extensible language modeling toolkit," in *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002*, 2002, pp. 901–904.

[35] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, California, USA: Association for Computational Linguistics, 1996, pp. 310–318.