SINET: Ethiop. J. Sci., 46(3): 237-249, 2023 ISSN: 0379–2897 (PRINT)
© College of Natural and Computational Sciences, Addis Ababa University, 2023 eISSN: 2520–7997

Date received: 08 January 2023; Date revised: 15 December 2023; Date accepted: 20 December 2023
DOI: https://dx.doi.org/10.4314/sinet.v46i3.2

DNN-based Multilingual Acoustic Modeling for Four Ethiopian Languages

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

[1]School of Information Science, Addis Ababa University, Ethiopia. E-mail: solomon.teferra@aau.edu.et
[2] Cognitive Systems Lab, University of Bremen

ABSTRACT: In this paper, we present the results of experiments conducted on multilingual acoustic modeling in the development of an Automatic Speech Recognition (ASR) system using speech data of phonetically much related Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta) with multilingual (ML) mix and multitask approaches. The use of speech data from only phonetically much related languages brought improvement over results reported in a previous work that used 26 languages (including the four languages). A maximum Word Error Rate (WER) reduction from 25.03% (in the previous work) to 21.52% has been achieved for Wolaytta, which is a relative WER reduction of 14.02%. As a result of using multilingual acoustic modeling for the development of an automatic speech recognition (ASR) system, a relative WER reduction of up to 7.36% (a WER reduction from 23.23% to 21.52%) has been achieved over a monolingual ASR. Compared to the ML mix, the multitask approach brought a better performance improvement (a relative WERs reduction of up to 5.9%). Experiments have also been conducted using Amharic and Tigrigna in a pair and Oromo and Wolaytta in another pair. The results of the experiments showed that languages with a relatively better language resources for lexical and language modeling (Amharic and Tigrigna) benefited from the use of speech data from only two languages. Generally, the findings show that the use of speech corpora of phonetically related languages with the multitask multilingual modeling approach for the development of ASR systems for less-resourced languages is a promising solution.

Keywords/ phrases: DNN, Multilingual Acoustic model, Under-resourced languages, Ethiopian Languages

INTRODUCTION

Automatic Speech Recognition (ASR) helps human to interact with different technologies in human languages. ASR enables automatic transcription of any speech, human-machine interaction via speech, assistive technologies, and speech translation. It increases social, political and economic development, by enabling people (especially illiterates and physically disabled) to use computing devices through speech in their own language.

Ethiopia has more than 80 languages and a population of about 120 million, with an illiteracy rate of about 49%. Consequently, ASR technologies in the Ethiopian languages are of high demand. However, due to lack of the required language resources, research attempts have been made for only a few of the more than 80 Ethiopian languages (Solomon, Menzel, and Tafila 2005; Munteanu et al. 2006; Solomon and Menzel 2007a, 2007b; Pellegrini and Lamel 2006, 2009; Martha, Solomon, and Menzel 2009; Martha 2010; Martha et al. 2012; Martha, Solomon, and Besacier 2014; Martha and Solomon 2015; Adey and Martha 2015; Gelas et al. 2011; Martha, Solomon, and Menzel 2011, 2010). Moreover, almost all the researchers have been challenged by the lack of speech and language resources.

Although there are some attempts towards the preparation of speech corpora for a few Ethiopian languages (Solomon, Menzel, and Tafila 2005; (Hafte and Sebisibe 2018; Solomon et al. 2020), the size of each of these corpora, however, is very small compared to speech corpora of other economically and technologically favored languages that have hundreds of hours of training speech. Moreover, the development of such speech corpora is not an easy and economically viable task to cover the more than 80 Ethiopian languages. On the contrary, to fully

*Author to whom correspondence should be addressed.

benefit from the modeling capacity of Deep Neural Networks (DNN), which have performed well in the development of acoustic models (AM) for ASR systems (Hinton et al. 2012; Gandhe, Metze, and Lane 2014; Shulby et al. 2017; Martha et al. 2020), we need much more training data. Otherwise, we will face the problem of overfitting or we need to stop learning very early as soon as the performance of our models start degrading on the held-out validation set. As it has been stated by (Hinton et al. 2012; Li et al. 2019), very large training sets can reduce overfitting while preserving modeling power.

As a solution for the above stated problems, Multilingual Automatic Speech Recognition (MLASR) has been suggested to develop an ASR system for an under-resourced language using existing training data of other languages (Heigold et al. 2013; Li et al. 2019; Solomon, Martha, and Schultz 2021; Martha, Solomon, and Schultz 2022; Weng et al. 1997). MLASR is described, in (Vu et al. 2014), as an ASR system for which one of the components (acoustic, language, or lexical model) is developed using training corpora in multiple languages.

The fact that almost all Ethiopian languages are under-resourced makes MLASR attractive for these languages. However, only few attempts (Martha et al. 2020; Martha, Solomon, and Schultz 2020c, 2020d; Solomon, Martha, and Schultz 2020; Martha, Solomon, and Schultz 2020b, 2022) have been made towards the development of MLASR for the Ethiopian languages, especially using the state of the art machine learning algorithms such as DNN. Although the previous works showed different appealing results and approaches, they did not investigate the use of phonetically much related languages in MLASR.

On the other hand, literature (Huang et al. 2013; Dalmia et al. 2018) show that in the development of MLASR systems, source languages which are phonetically related to the target language help more than the phonetically distant ones. A previous work (Martha, Solomon, and Schultz 2020a), that analyzed phonetic relationship among and between Ethiopian and GlobalPhone (a speech database of 22 languages) (Schultz, Vu, and Schlippe 2013) languages, revealed that four Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta) are related to each other more than the relation they have with the other GlobalPhone languages. The study also showed that very high phonetic overlap may exist among languages that belong to different language groups. For example, although Oromo and Wolaytta are from different language groups, Oromo phone set covers 97.3% of Wolaytta phones while 92.3% of Oromo phones are included in Wolaytta phone set. The analysis also showed that the highest phonetic overlap is seen between languages that are in the same language group which are Amharic and Tigrigna. It revealed that Amharic phones are fully (100%) covered by the Tigrigna phone set while about 90% of the Tigrigna phones are covered in the Amharic phone set.

In this paper, we present the results of experiments conducted on the development of MLASR system using only speech corpora of the four phonetically much related Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta for multilingual acoustic modeling. In the experiments, we have compared the ML mix and multitask MLASR development approaches. The use of speech data of two pairs of languages (Amharic and Tigrigna in one while Oromo and Wolaytta in another pair) with the highest phonetic similarity has also been investigated.

The paper is organized as follows. In the first 2 sections, we give the introduction and motivations of our work and present a brief review of literature on the development of MLASR, respectively. The description of phonetic and morphological features of the considered Ethiopian languages is presented in the third section. The next 2 consecutive sections present the corpora used in our experiments and the experimental setup, respectively. In the last 3 sections we presented the results of all our experiments on the development of multilingual acoustic modeling, discussions of the results and the conclusions drawn from our findings, respectively.

Multilingual ASR

As indicated in the introduction section, when language resources from multiple languages are used to develop one or more of the components of an ASR system (acoustic, language and/or lexical models), the resulting ASR system becomes a Multilingual one.

MLASR systems are appealing solutions for under-resourced languages in which training speech corpora are sparse or do not exist at all (Schultz and Waibel 2001). Furthermore, they are helpful for multilingual, multi-ethnic, and economically disadvantaged countries like Ethiopia.

MLASR has been investigated using different approaches such as Gaussian Mixture Modeling and recently, Deep Neural Network (DNN) models. Currently, the application of DNNs is resulting in performance improvement for MLASR systems (Heigold et al. 2013; Li et al. 2019; Solomon, Martha, and Schultz 2021; Martha, Solomon, and Schultz 2022).

There are, however, three major factors that affect the performance of DNN-based acoustic modeling for a MLASR system: the amount of training data we get from source languages, the amount of training data we have for the target language and the linguistic distance between the source languages and the target language. Literature show that using various source languages increases the chance of having more generalized multilingual DNN with better context coverage. On the other hand, the difference between the target language and the source language(s) may obtrude with impurification of training data and hurt target language's acoustic model (Lin et al. 2009; Vu et al. 2014). Especially, when reasonable amount of training data for target language is available, the negative effect of language mismatch may even make MLASR system perform worse than the monolingual system. It is shown that the MLASR trained on the similar language(s) outperforms the one trained on all available source languages. Furthermore, a set of experiments are provided to investigate whether it is better to utilize data from similar languages or more data from diverse languages in the MLASR training (Müller et al. 2014). It is shown that when MLASR training employs "best fitting" languages, significant improvement is obtained. It has been shown that adding mismatched languages gives gains over the monolingual baseline if the set of source languages is big enough to train a robust model. Different researches have been conducted on the use of different sets of source languages for the development of MLASR for Ethiopian languages. One of these works is (Martha, Solomon, and Schultz 2020c) that has conducted two sets of experiments on the development of MLASR: 1) The use of the training speech of the target language itself with the 22 GlobalPhone corpora. 2) The use of the training speech of the target corpora with the 25 mixed (GlobalPhone and 3 corpora of the Ethiopian languages) source languages. In this work only the Multilingual (ML) mix and weight transfer/adaptation approaches have been applied.

The recently published work of the same authors (Martha, Solomon, and Schultz 2022) presented different investigations of MLASR for 26 languages including the four Ethiopian languages considering different degrees of phonetic relatedness of the languages using the ML mix, transfer and multitask approaches. However, although the four Ethiopian languages are more phonetically related with each other, investigation on only these languages in MLASR was not conducted. In the current work, investigation has been conducted using only the four phonetically much related Ethiopian languages for the acoustic model component of an ASR system. Moreover, experiments have also been conducted using only two of the most phonetically related languages.

DNN-based Multilingual ASR

Although Artificial Neural Networks (ANNs) have been introduced in the area of ASR in the 1940s, they did not outperform the Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) until 2009. Since 2009, Deep Neural Networks (DNN) have become very popular in ASR for the hybrid HMM-DNN systems outperformed the dominant HMM-GMM on the same data (Hinton et al. 2012). These developments in the application of DNN in ASR research brought more achievements in the development of MLASR than the achievement gained in the development of monolingual ASR models. Models capable of learning from multiple languages have been studied using hybrid HMM-DNN (Heigold et al. 2013; Huang et al. 2013; Markus Müller, Stüker, and Waibel 2016; Dalmia et al. 2018; Li et al. 2019; Martha, Solomon, and Schultz 2022). Different DNN architectures are used in the development of ASR systems. One of the architectures is Time Delay Neural Networks (TDNNs). TDNNs architectures are efficient and achieve better performance for ASR (Peddinti, Povey, and Khudanpur 2015) for they are able to learn long term temporal contexts. Moreover, by using singular value decomposition (SVD), the number of parameters in TDNN models is reduced making them less expensive than the Recurrent Neural Networks (RNN). The factored form of TDNNs (TDNNf) (Povey et al. 2018) is similar with TDNN in its structure, but it is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal. TDNNf achieved better performance and effectiveness than TDNN in under-resourced scenarios. We have used this DNN architecture in all of our experiments for the development of MLASR systems.

Researchers have been experimenting on the use of different approaches for the development of DNN-based MLASR. Multilingual mix (ML mix) (Sailor and Hain 2020; Fathima et al. 2018; Hara and Nishizaki 2017), and weight transfer or multilingual adaptation (Liu et al. 2018; Tong, Garner, and Bourlard 2017) and multitask (Heigold et al. 2013; Huang et al. 2013) are the common ones. Since ML mix and multitask are used in our experiment, we present a brief description of the two as follows.

In ML mix, all the training resources (lexicon, audio data and their transcription) of the source languages are combined to make one training lexicon with a universal phone set and training speech corpus (audio and transcription). The combined resource is used to train one acoustic model. The tied-states for training the multilingual DNN AMs are obtained by using the multilingual GMM-HMM systems to build multilingual decision trees and generate tied-state alignments. So there is no language information in the AMs. The universal AM is used in decoding the target language using the language specific language model (LM) and decoding lexicon of the target language. If there are language-specific phones, which are not covered by the universal training phone set, they will be mapped to the nearest phone in the universal phone set.

Multitask modeling is learning multiple tasks in parallel and use a shared representation (Heigold et al. 2013). It is adopted from the architectures developed to solve the problem of making a robust AM to be tuned for different domains and/or noise levels.

In MLASR, each language is considered as a task. Multitask approach has enabled the development of an AM with same hidden layer and language-specific output (softmax) layers. In this approach, all the training data (from all the languages) are used to train the hidden layers and the language specific training data is used to train the softmax layers.

Ethiopian Languages

Ethiopia is one of the multilingual and multi-ethnic countries in which more than 80 languages are spoken. Ethnologue[1] states that, "The number of individual languages listed for Ethiopia is 90. Of these, 88 are living and 2 are extinct. Of the living languages, 85 are indigenous and 3 are nonindigenous. Ethiopian languages belong to four major language groups: Semitic, Cushitic, Omotic and Nilo-Saharan."

The languages considered in this work are Amharic and Tigrigna that belong to the Semitic, Oromo from the Cushitic and Wolaytta from the Omotic language groups. All of these language groups fall under Afro-Asiatic language family. Based on the 2021 data on Ethnologue[2], there are more than 57.4 and 9.8 million people who speak Amharic and Tigrigna, respectively while more than 37 and 2.5 million people speak Oromo and Wolaytta, respectively.

These languages are used for different communication purposes in Ethiopia. Amharic serves as the working language of the Federal Government and the Amhara and other regional states. The Tigray and Oromiya regional states use Tigrigna and Oromo as their working languages, respectively. Several websites and other electronic media like news, blogs and social media are being developed in these languages. The languages also serve as medium of instructions in primary and secondary schools. Google also offers a searching capability in Amharic, Tigrigna and Oromo. Furthermore, Google also developed Amharic translation system that is released for public use. Since three out of the four

[1] https://www.ethnologue.com/browse/countries

[2] https://www.ethnologue.com/browse/names

language groups in Ethiopia are considered in our work, we believe that the concept proofed for these languages can be applied for the other Ethiopian languages.

Phonology

Even if these four languages belong to three different language groups, they share about 70% of their phone sets, including the ejectives: t′ k′ p′ ts′ tʃ′. In this subsection, we describe only a few phonetic relations between two language pairs: Amharic-Tigrigna and Oromo-Wolaytta. 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. The four Tigrigna sounds that are not found in Amharic are ʕ, ħ, x and x́. 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 glottalized or ejective: t′ k′ p′ ts′ tʃ′ sounds are found in both Amharic and Tigrigna (Leslau, 2000). Consonant gemination brings semantic difference in these languages. Both languages use 7 vowels: ə, u, i, a, e, ɨ, o.

Although Oromo belongs to the Cushitic and Wolaytta to the Omotic language group, they have more phonetic overlap than the overlap they have with languages in the other pair. Each of them have five vowels that have long and short variants. This makes up the vowel set of each language to be ten. Both of them are also tonal languages. Oromo has 28 consonants and Wolaytta 27. These languages share a number of consonants except the Oromo consonants ɲ and x are not used in Wolaytta while the Wolaytta consonant ʒ is not used in Oromo.

Morphology

Although these four languages have rich morphology that use nominals and verbs that are inflected for person, number, gender, tense, aspect, and mood, (Griefenow-Mewis 2001), we can categorize Amharic and Tigrigna to one pair while Oromo and Wolaytta to another. Amharic (Wolf 2000) and Tigrigna (Tewolde 2002), use root-pattern morphology. This pair has more morphological complexity than the pair of Oromo and Wolaytta. Unlike the Semitic languages, Oromo and Wolaytta are suffixing languages. The difference in their morphological complexity has been observed in the higher Out of Vocabulary (OOV) rate of Amharic and Tigrigna pair than the Oromo and Wolaytta using the training vocabulary (word type of the training speech transcription) as presented in Table 1. The last column of the Table shows OOVs on the same vocabulary size (21,232) for all the languages.

Table 1. OOV of the four Ethiopian Languages.

Languages	Training Vocabulary	OOV	OOV with 21,232
Amharic (AMH)	28,661	24.99	33.37
Tigrigna (TIR)	31,759	16.33	19.75
Oromo (ORM)	21,232	11.73	11.73
Wolaytta (WAL)	25,267	9.34	10.09

Writing System

Amharic and Tigrigna use Ethiopic while Oromo and Wolaytta use the Latin scripts for writing. The Ethiopic script is a syllabic script where each character is the representation of a consonant and a vowel. This writing system does not show consonant gemination and presence or absence of the epenthetic vowel and the glottal stop consonant. In contrast, the current writers in Oromo and Wolaytta write the geminated and the non-geminated consonants as double letters and single letter, respectively. They also show the long and short vowels in their writing. Short vowels are represented by single letters whereas the long ones are represented by double letters. Since the scripts used in all the languages have relatively clear and consistent grapheme-to-phoneme (G2P) relations, we have generated the pronunciation dictionaries required for the development of the MLASR system automatically.

Speech and Text Corpora

Speech Corpora

The speech corpora we have used consist of a read speech corpus developed for Amharic by (Solomon, Menzel, and Tafila 2005), and the newly developed speech corpora of the four Ethiopian languages (Solomon et al. 2020). That means we have used five corpora of the four languages (two separately prepared corpora for Amharic). There is no special reason for using two corpora for Amharic except its availability. In this subsection, we give a brief description of all of them. For more details, we direct readers to the original publications.

The Amharic corpus is a read speech corpus prepared as part of a PhD research project conducted at the University of Hamburg (Solomon, Menzel, and Tafila 2005). It has 20 hours of training speech and 100 training readers who read a total of 10,850 sentences (28,666 tokens), development and evaluation sets recorded from 20 speakers (10 each). The corpus has 5000 and 20000 development sets as well as 5000 and 20000 evaluations sets. In this experiment, we have merged the development sets and evaluations sets so as to evaluate the ASR systems with relatively bigger (in size) development and evaluation sets. The Amharic corpus consists of development and evaluation test sets of 760 utterances with about 1.5 hours of speech, each. AMH2005 stands for this corpus.

The other corpora are the ones developed for Amharic, Tigrigna, Oromo and Wolaytta by a thematic research funded by the Addis Ababa University (Solomon et al. 2020). From the total recordings, four speakers have been held out for development and evaluation sets, each. In the selection of the test sets gender balance has been considered. Table 2 summarizes the amount of speech data in each set for the four corpora developed for the Ethiopian languages. AMH2020 stands for the Amharic corpus prepared in the thematic research.

Table 2: Details on Corpora of the Ethiopian Languages.

Sets of Corpora	Units	Corpora				
		AMH2005	AMH2020	TIR	ORM	WAL
Training	Speech size in hours	20	24	22.1	22.8	29.7
	No of Speakers	100	90	90	90	77
	No of Utterances	10,875	11,274	11,305	11,297	10,939
Development	Speech size in hours	1.5	1.2	1.1	1.2	1.5
	No of Speakers	10	4	4	4	4
	No of Utterances	760	507	511	505	553
Evaluation	Speech size in hours	1.5	1.3	1	1.1	1.7
	No of Speakers	10	4	4	4	4
	No of Utterances	760	508	507	501	578

Text Corpora

The text corpus used for the development of the LMs for these languages in the previous researches have been used in this work. We could get access to a relatively bigger text corpus for Amharic and Tigrigna (about 4 Million word tokens, each). We have used a few text corpus, which is a mix of text from different domains including spiritual domain and made available online for Oromo (Suchomel and Rychlý 2016). To minimize the negative effect of out of domain text, we were required to select only a part (1.5 Million tokens) of the text with minimum domain difference from the transcriptions of our speech corpus. To this end, we used sentence based perplexities computed from a 9-gram character LM developed using the transcriptions of the training speech. For Wolaytta, which has less presence on the web, we could not get any text corpus and, therefore, we used only the transcription of the speech corpus for language modeling.

Experimental setup for the multilingual ASR systems

All the multilingual AMs were built using Kaldi ASR toolkit (Povey et al. 2011). First we built context dependent HMM-GMM based AMs using 39 dimensional mel-frequency cepstral coefficients (MFCCs) to each of which cepstral mean and variance normalization (CMVN) have been applied. The AM uses a fully continuous HMM with 3 emitting states moving from left to right. Then feature transformation has been done using Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) for each of the models. Finally, Speaker Adaptive Training (SAT) has been applied using an affine transform, and feature space Maximum Likelihood Linear Regression (fMLLR). Among all the AMs, the best model is used to obtain alignments for DNN training.

In DNN acoustic modeling, the same speech data that is used to train HMM-GMM models has been used. However, three-fold data augmentation (Ko et al. 2015) has been applied before the extraction of 40-dimensional MFCCs without derivatives. We have also extracted 3-dimensional pitch features and 100-dimensional i-vectors for the purpose of speaker adaptation. The architecture we used is Factored Time Delay Neural Networks (Povey et al. 2018) with additional Convolutional layers (CNN-TDNNf) that is adopted from the standard Kaldi WSJ recipe. Our network has 6 CNN layers followed by 9 TDNNf layers and one rank reduction layer. The TDNNf consists of 1024 units and 128 bottleneck units. But for the TDNNf layer immediately following the CNN layers we have increased the number of the bottleneck units to 256. The default hyper-parameters of the standard recipe were used. The same DNN architecture is used in the development of ML AMs using the two approaches: ML mix and multitask.

In the development of AMs with the ML mix approach, the training speech, the transcription, the training Pronunciation Dictionaries (PDs) and the phone sets of all the involved languages are mixed and used as a single training resource to train the ML AMs. Therefore, there is no language information at any of the DNN layers. We have adapted the WSJ recipe for the development of ML AMs using ML mix approach.

In the multitask approach, each language is considered as a task. The data from all the involved languages is used to train the hidden layers of the neural network while the output layer is specific to each language. For the development of the AMs with the multitask approach, we have adapted the recently provided multitask recipe for chain models from babel multilingual example by modifying the feature extraction and the DNN architecture.

All the AMs in this work are evaluated on the test set of the respective language using the respective decoding PD and its monolingual Language Model (LM). We have used decoding PDs and trigram LMs presented in Table 3 for the evaluation of all the monolingual and ML ASR systems.

Table 3. PDs and LMs used for evaluating all the AMs.

Corpora	PD size in thousands	OOV in %	LM perplexity
AMH2005	310	3.06	41.2
AMH2020	323	6.21	241.26
TIR	299	4.89	172.42
ORM	21.23	11.73	266.17
WAL	25.27	9.34	254.9

Multilingual Acoustic Modeling for Ethiopian Languages

As it has been stated in the introduction section, we have learned from literature that the benefit we gain from DNN-based acoustic modeling for a MLASR depends on the phonetic relation among the languages, the amount of training data we have in the target and source languages and the DNN approach we apply. We have, therefore, conducted several experiments towards the development of MLASR using two ML approaches and two levels of phonetic relations among the languages considered in our study. The results of the experiments are presented in the next 2 subsections.

For our first set of experiments, we considered the five speech corpora we have in the four Ethiopian languages. In these experiments we have compared the performance of ML AMs developed using the ML mix and the ones developed using multitask DNN approaches. The results are presented in the first subsection of this section.

For the second set of experiments we used speech data from only two languages in the ML AM training. In these sets of experiments, we paired the Ethiopian languages into two based on their phonetic relatedness. Amharic and Tigrigna in one pair and Oromo and Wolaytta in another. We have presented the results of this set of experiments in the second subsection of this section.

Multilingual Acoustic Modeling Using Four Ethiopian Languages

In this subsection we present the results of our experiments that compare the performance of the two DNN approaches using five speech corpora in four Ethiopian languages. The results of our experiments with the ML mix approach are presented in Table 4. The ML26 in the table stands for the ML AMs developed using training data of 26 languages presented in a previous work (Martha, Solomon, and Schultz 2022) while ML4 stands for the ML AMs that are developed using only the 5 training speech corpora we have for the four Ethiopian languages. The monolingual ASR WERs presented in Table 4 are also from (Martha, Solomon, and Schultz 2022). As we can see from the table, reducing the source languages from 26 that includes phonetically distant languages from the GP to only the much related 4 Ethiopian languages resulted in WER reduction for all the languages. We have also presented the relative WER reductions resulted from limiting source corpora to only phonetically much related languages in the last column of Table 4.

Table 4: Performance of Multilingual AMs ML4 using ML mix.

Languages/ Corpora	WERs			Relative WER Reduction of ML4 over ML26
	Monolingual	ML26	ML4	
AMH2005	8.43	8.45	8.25	2.37
AMH2020	18.88	20.34	19.35	4.87
TIR	16.82	18.39	17.24	6.25
ORM	32.28	33.74	32.37	4.06
WAL	23.23	24.56	22.87	6.88

Although we have got better MLASR by excluding distant languages, we did not benefit from developing MLASR systems over the monolingual ones, except for AMH2005 and WAL that got relative WER reduction of 2.14\% and 1.55\%, respectively. This is due to the fact that in ML mix approach the resources of all the languages are combined to develop one general AM that does not have specific language information. So we have experimented with multitask approach. For the multitask approach, each of the five corpora is considered as a task of different nature. We have conducted several experiments to see the effect of number of epochs and data weights on the performance of the system. Although we could not get any specific epoch that worked the same way for all the corpora, we have taken the results that are optimal for most of the tasks that is with epoch 6.

The performance of the MLASR systems that use the set of the ML AMs developed with multitask approach is presented in Table 5. For comparison purpose, we also presented the performances of ML26 developed using the multitask approach and the monolingual ASR presented in (Martha, Solomon, and Schultz 2022) as well as ML4 developed using ML mix approach in the Table. As we can see in the Table, the approaches bring difference in the performance of MLASR. The ML AMs trained using multitask approach brought up to 5.90% (for WAL) relative WER reductions over the ML AMs that are trained using the ML mix approach. The use of multitask approach also brought up to 7.36\% (for WAL) relative WER reduction over the monolingual AMs.

Moreover, the use of only much phonetically related languages in MLASR using multitask approach brought improvement over the use of phonetically distant languages using the same approach. As shown in Table 5, ML4 developed using multitask approach resulted in a relative WER reduction of 14.02 over the ML26 developed using the same approach.

Table 5. Performance of Multilingual AMs with Multitask approaches.

Languages/ Corpora	WERs				Relative WER Reduction of ML4 with Multitask Over		
	Monolingual	ML 26 Multitask	ML4		Monolingual	ML26 Multitask	ML mix ML4
			ML mix	Multitask			
AMH2005	8.43	8.24	8.25	8.16	3.20	0.97	1.09
AMH2020	18.88	19.64	19.35	18.95	-0.37	3.51	2.07
TIR	16.82	17.21	17.24	16.77	0.30	2.56	2.73
ORM	32.28	33.04	32.37	31.73	1.70	3.96	1.98
WAL	23.23	25.03	22.87	21.52	7.36	14.02	5.90

MLASR using Two Highly Related Languages

As we can see from the results presented in the previous subsection, reducing the number of languages to only Ethiopian languages resulted in performance improvement over the ML26 MLASR systems presented in (Martha, Solomon, and Schultz 2022). We have, therefore, further reduced the number of languages to two phonetically much related languages and developed MLASR systems using both ML mix and multitask approaches. For this experiment, we have considered Amharic and Tigrigna as one pair and Oromo and Wolaytta as another due to their phonetic relation. As we did in the previous sets of experiments in the use of multitask approach, we have applied different number of epochs to get an optimal one. We found out that using epoch 7 is better for AMH2005 and TIR pair while using epoch 6 is better for ORM and WAL pair. We have presented the results in Table 6.

Table 6: Performance of MLASRs with the Highest Phonetic Relatedness.

Languages/Corpora	WERs					Relative WER Reduction of ML2 over ML4	
	Monolingual	ML Mix		Multitask			
		ML4	ML2	ML4	ML2	ML mix	Multitask
AMH2005	8.43	8.25	8.08	8.16	8.01	2.06	1.84
TIR	16.82	17.24	16.66	16.77	16.58	3.36	1.13
ORM	32.28	32.37	31.78	31.73	32.21	1.82	-1.51
WAL	23.23	22.87	23.45	21.52	21.92	-2.54	-1.86

The results presented in Table 6 show that using only two phonetically much related languages in ML AM developed using ML mix approach brought improvement in performance for all languages, but Wolaytta, over the ML4 model that is developed using speech data of the four Ethiopian languages. From the results, we can observe that the use of ML mix approach for the development of MLASR systems using only phonetically related languages leads to performance improvement instead of using resources of phonetically distant languages in MLASR.

When multitask approach is used Amharic and Tigrigna have got performance improvement over the ML4 while Oromo and Wolaytta did not. This can be attributed to the quantity and quality of the language specific resources (pronunciation dictionary and language model) used during decoding. In this regard, Amharic and Tigrigna used relatively large pronunciation dictionary and good language model. However, due to lack of resources, this could not be the case in Oromo and Wolaytta. Thus for Oromo and Wolaytta, the acoustic model developed with more data (all Ethiopian languages' corpora) seem to be stronger than the model developed using only the corpora of the two languages.

We have also compared the WER reduction gained from the use of only much related language pairs over the WER of the monolingual model that is developed using only the training speech of the target language. The results are presented in Table 7. In this set of experiments the multitask approach brought a higher WER reduction over the monolingual models than the ML mix approach, except for Oromo.

Table 7. Performance of ML2 Vs. the monolingual AMs.

Languages/ Corpora	WERs			Relative WER Reduction of ML2 Over Monolingual	
	Monolingual	ML mix	multitask	ML mix	Multitask
AMH2005	8.43	8.08	8.01	4.15	4.98
TIR	16.82	16.66	16.58	0.95	1.43
ORM	32.28	31.78	32.21	1.55	0.22
WAL	23.23	23.45	21.92	-0.95	5.64

DISCUSSIONS OF RESULTS

In this study we have investigated the use of speech data from phonetically much related languages in the development of MLASR. Our results show that the use of speech corpora of phonetically much related languages in MLASR training brings performance improvement. The comparison of our results with the results presented in (Martha, Solomon and Schultz, 2022) confirmed this. In (Martha, Solomon and Schultz, 2022), speech data from 26 languages (including the four Ethiopian languages) have been used to develop MLASR system. In both the approaches we have used (ML mix and multitask), the ML4 MLASR systems developed in our work have lower WER than the ML26 MLASR systems developed using 26 languages. In (Martha, Solomon and Schultz, 2022), MLASR experiments were conducted using speech data from 10 and 14 phonetically related languages, as well as 10 related and unrelated languages using only the multitask approach. Our ML4 MLASR systems developed using multitask approach far better than almost all of the systems developed in (Martha, Solomon and Schultz, 2022).

Considering the most related languages has also an advantage for languages that have relatively better language resources for language and lexical modeling. This is vivid, when we see the results we have got in ML2 MLASR system. Both Amharic and Tigrigna have got improvement in ML2 over ML4. But this is not the case for Oromo and Wolaytta, both of which used smaller vocabulary lexical model and smaller amount of text data for language model training than the former pair (Amharic and Tigrigna).

Compared to ML mix, in all our experiments the multitask approach leads to better performance. This is also true in the MLASR results presented in (Martha, Solomon and Schultz, 2022).

CONCLUSIONS

In this paper, we have presented the results of our experiments conducted towards the development of MLASR using five speech corpora in four Ethiopian languages. We have used the different monolingual and multilingual acoustic models presented in previous work (Martha, Solomon, and Schultz 2022) as baseline systems against which we measure the benefit we get from the current experiments.

The results of our experiments show that phonetic relationship among languages is a factor for the performance improvement of an MLASR system. The more they are related, the lower the WERs we achieve.

The results of our experiments also showed that the approach we choose for the development has also a significant impact on the performance of an MLASR. We have observed that the multitask approach has outperformed the ML mix approach in all the cases, except for Oromo when we apply it for the Oromo-Wolaytta language pairs.

Generally, our research confirmed that the use of MLASR outperforms the monolingual ASR in acoustic modeling especially for related languages. So one can extend the coverage of MLASR development for a lot of new Ethiopian languages with a minimum investment on the development of training speech corpora and by using the five existing speech corpora in the four Ethiopian languages.

ACKNOWLEDGMENT

We are grateful to the Alexander von Humboldt Foundation that financed the research stay of the first two authors as experienced researchers at the Cognitive Systems Lab (CSL) of the University of Bremen. Some of the experiments have been conducted during the research stay. The foundation also continued the financial support for the two authors with its return fellowship program, during which a number of additional experiments have been conducted in the same lab.

REFERENCES

1. Adey, Edessa Dribssa, and Martha Yifiru Tachbelie. 2015. "Investigating the Use of Syllable Acoustic Units for Amharic Speech Recognition." In *{AFRICON} 2015, Addis Ababa, Ethiopia, September 14-17, 2015*, 1–5.https://doi.org/10.1109/AFRCON.2015.7331999.
2. Dalmia, Siddharth, R Sanabria, F Metze, and A Black. 2018. "Sequence-Based Multi-Lingual Low Resource Speech Recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4909–13.
3. Fathima, N, Tanvina Patel, C Mahima, and Anuroop Iyengar. 2018. "TDNN-Based Multilingual Speech Recognition System for Low Resource Indian Languages." In *INTERSPEECH*.
4. Gandhe, A, F Metze, and I Lane. 2014. "Neural Network Language Models for Low Resource Languages." In *INTERSPEECH*.
5. Gelas, Hadrien, Solomon Teferra Abate, Laurent Besacier, and François Pellegrino. 2011. "Quality Assessment of Crowdsourcing Transcriptions for African Languages." In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3065–68. https://doi.org/10.21437/interspeech.2011-767.
6. Griefenow-Mewis, Catherine. 2001. *A Grammatical Sketch of Written Oromo*.
7. Hafte, Abera, and Sebsibe Hailemariam. 2018. "Design of a {T}igrinya Language Speech Corpus for Speech Recognition." In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, 78–82. Santa Fe, New Mexico, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-3811.
8. Hara, S, and H Nishizaki. 2017. "Acoustic Modeling with a Shared Phoneme Set for Multilingual Speech Recognition without Code-Switching." In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1617–20. https://doi.org/10.1109/APSIPA.2017.8282284.
9. Heigold, G, V Vanhoucke, A Senior, P Nguyen, M Ranzato, M Devin, and J Dean. 2013. "Multilingual Acoustic Models Using Distributed Deep Neural Networks." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8619–23.
10. Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29 (6): 82–97.
11. Huang, J, J Li, D Yu, L Deng, and Y Gong. 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*, 7304–8.
12. Ko, Tom, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. "Audio Augmentation for Speech Recognition." In *INTERSPEECH*.
13. Li, Xinjian, Siddharth Dalmia, Alan Black, and Florian Metze. 2019. "Multilingual Speech Recognition with Corpus Relatedness Sampling." *Interspeech 2019*.
14. Lin, Hui, Li Deng, Dong Yu, Yifan Gong, Alex Acero, and Chin-Hui Lee. 2009. "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR." *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4333–36.
15. Liu, D, X Wan, J Xu, and P Zhang. 2018. "Multilingual Speech Recognition Training and Adaptation with Language-Specific Gate Units." In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 86–90.

https://doi.org/10.1109/ISCSLP.2018.8706584.

16. Martha, Yifiru Tachbelie. 2010. "Morphology-Based Language Modeling for Amharic." University of Hamburg. http://www.sub.uni-hamburg.de/opus/volltexte/2010/4848/index.html.

17. Martha, Yifiru Tachbelie, and Solomon Teferra Abate. 2015. "Effect of Language Resources on Automatic Speech Recognition for Amharic." In *IEEE AFRICON Conference*. Vol. 2015-Novem. https://doi.org/10.1109/AFRCON.2015.7331871.

18. Martha, Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2014. "Using Different Acoustic, Lexical and Language Modeling Units for ASR of an under-Resourced Language - Amharic." *Speech Communication* 56 (1): 181–94. https://doi.org/10.1016/j.specom.2013.01.008.

19. Martha, Yifiru Tachbelie, Solomon Teferra Abate, Laurent Besacier, and Solange Rossato. 2012. "Syllable-Based and Hybrid Acoustic Models for Amharic Speech Recognition." In *Third Workshop on {SLTU}, Cape Town, South Africa, May 7-9, 2012*, 5–10. http://www.isca-speech.org/archive/sltu_2012/su12_005.html

20. Martha, Yifiru Tachbelie, Solomon Teferra Abate, and Wolfgang Menzel. 2009. "Morpheme-Based and Factored Language Modeling for Amharic Speech Recognition." In *Human Language Technology. Challenges for Computer Science and Linguistics - 4th Language and Technology Conference, {LTC} 2009, Poznan, Poland, November 6-8, 2009, Revised Selected Papers*, 82–93. https://doi.org/10.1007/978-3-642-20095-3_8.

21. ———. 2010. "Morpheme-Based Automatic Speech Recognition for a Morphologically Rich Language - Amharic." In *2nd Workshop on Spoken Language Technologies for Under-Resourced Languages, {SLTU} 2010, Penang, Malaysia, May 3-5, 2010*, 68–73. http://www.isca-speech.org/archive/SLTU_2010/su10_068.html.

22. ———. 2011. "Morpheme-Based and Factored Language Modeling for Amharic Speech Recognition." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6562 LNAI:82–93. https://doi.org/10.1007/978-3-642-20095-3_8.

23. Martha, Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020a. "Analysis of GlobalPhone and Ethiopian Languages Speech Corpora for Multilingual ASR." In *LREC 2020*.

24. ———. 2020b. "Deep Neural Networks Based Automatic Speech Recognition For Four Ethiopian Languages." In *ICASSP 2020*.

25. ———. 2020c. "Development of Multilingual ASR Using GlobalPhone for Less-Resourced Languages: The Case of Ethiopian Languages." In *Proc. Interspeech 2020*, 1032–36. https://doi.org/10.21437/Interspeech.2020-2827.

26. ———. 2020d. "DNN-Based Multilingual Automatic Speech Recognition for Wolaytta Using Oromo Speech." In *SLTU/CCURL@LREC*.

27. ———. 2022. "Multilingual Speech Recognition for GlobalPhone Languages." *Speech Communication* 140: 71–86. https://doi.org/https://doi.org/10.1016/j.specom.2022.03.006.

28. Martha, Yifiru Tachbelie, Solomon Teferra Abate, Tanja Schultz, and Ayimunishagu Abulimiti. 2020. "Multilingual Speech Recognition for GlobalPhone Languages." In *LREC 2020*.

29. Müller, M, S Stüker, Zaid Sheikh, F Metze, and A Waibel. 2014. "MULTILINGUAL DEEP BOTTLE NECK FEATURES A STUDY ON LANGUAGE SELECTION AND TRAINING TECHNIQUES." In .

30. Müller, Markus, Sebastian Stüker, and Alex Waibel. 2016. "Language Adaptive DNNs for Improved Low Resource Speech Recognition." In *Interspeech 2016*, 3878–82. https://doi.org/10.21437/Interspeech.2016-1143.

31. Munteanu, Cosmin, Gerald Penn, Ron Baecker, and Yuecheng Zhang. 2006. "Automatic Speech Recognition for Webcasts." University of Hamburg. https://doi.org/10.1145/1180995.1181005.

32. Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur. 2015. "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts." In *Sixteenth Annual Conference of the International Speech Communication Association*.

33. Pellegrini, Thomas, and Lori Lamel. 2006. "Experimental Detection of Vowel Pronunciation Variants in Amharic." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, {LREC} 2006, Genoa, Italy, May 22-28, 2006.*, 1005–8. http://www.lrec-conf.org/proceedings/lrec2006/summaries/701.html.

34. ———. 2009. "Automatic Word Decompounding for {ASR} in a Morphologically Rich Language: Application to Amharic." *{IEEE} Trans. Audio, Speech & Language Processing* 17 (5): 863–73. https://doi.org/10.1109/TASL.2009.2022295.

35. Povey, Daniel, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks." In *Interspeech*, 3743–47.

36. Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, et al. 2011. "The Kaldi Speech Recognition Toolkit." In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

37. Sailor, Hardik B, and Thomas Hain. 2020. "Multilingual Speech Recognition Using Language-Specific Phoneme Recognition as Auxiliary Task for Indian Languages." In *INTERSPEECH*.

38. Schultz, Tanja, Ngoc Thang Vu, and Tim Schlippe. 2013. "GlobalPhone: A Multilingual Text Amp; Speech Database in 20 Languages." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8126–30. https://doi.org/10.1109/ICASSP.2013.6639248.

39. Schultz, Tanja, and Alex Waibel. 2001. "Language-Independent and Language-Adaptive Acoustic Modeling for Speech Recognition." *Speech Commun.* 35 (1–2): 31–51. https://doi.org/10.1016/S0167-6393(00)00094-7.

40. Shulby, Christopher Dane, Martha Dais Ferreira, Rodrigo Fernandes de Mello, and Sandra M Aluísio. 2017. "Acoustic Modeling Using a Shallow CNN-HTSVM Architecture." *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, 85–90.

41. Solomon, Teferra Abate, and Wolfgang Menzel. 2007a. "Automatic Speech Recognition for an Under-Resourced Language - Amharic." In .2007b. "Syllable-Based Speech Recognition for Amharic." In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, SEMITIC@ACL 2007, Prague, Czech Republic, June 28, 2007*, 33–40. https://www.aclweb.org/anthology/W07-0805/.

42. Solomon, Teferra Abate, Wolfgang Menzel, and Bairu Tafila. 2005. "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition." In *9th European Conference on Speech Communication and Technology*, 1601–4. https://doi.org/10.21437/interspeech.2005-467.

43. Solomon, Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Atinafu, and Biniyam Ephrem. 2020. "Large Vocabulary Read Speech Corpora for Four Ethiopian Languages: Amharic, Tigrigna, Oromo and Wolaytta." In *LREC 2020*.

44. Solomon, Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. "Multilingual Acoustic and Language Modeling for Ethio-Semitic Languages." In *Interspeech*, edited by Helen Meng, Bo Xu, and Thomas Fang Zheng, 1047–51. ISCA.

45. ———. 2021. "End-To-End Multilingual Automatic Speech Recognition for Less-Resourced Languages: The Case of Four Ethiopian Languages." In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP39728.2021.9415020.

46. Suchomel, V\'\it, and Pavel Rychlý. 2016. "Oromo Web Corpus." http://hdl.handle.net/11234/1-2588.

47. Tewolde, Yohannes Tesfay. 2002. *A Modern Grammar of Tigrinya*. Rome: Tipografia U. Detti.

48. Tong, S, Philip N Garner, and H Bourlard. 2017. "Multilingual Training and Cross-Lingual Adaptation on CTC-Based Acoustic Model." *ArXiv* abs/1711.1.

49. Vu, Ngoc Thang, David Imseng, Daniel Povey, Petr Motlícek, Tanja Schultz, and Hervé Bourlard. 2014. "Multilingual Deep Neural Network Based Acoustic Modeling for Rapid Language Adaptation." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7639–43.

50. Weng, Fuliang, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke. 1997. "A Study of Multilingual Speech Recognition." *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*.

51. Wolf, Leslau. 2000. Introductory Grammar of Amharic. Introductory Grammar of Amharic. Porta Linguarum Orientalium, Neue Serie, Bd. 21. Wiesbaden: Harrassowitz.