Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 61–69
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic

Yonas Woldemariam
Dept. Computing Science, Umeå University, Sweden
yonasd@cs.umu.se

Abstract

While building automatic speech recognition (ASR) requires a large amount of speech and text data, the problem gets worse for less-resourced languages. In this paper, we investigate a model adaptation method, namely transfer learning for a less-resourced Semitic language i.e., Amharic, to solve resource scarcity problems in speech recognition development and improve the Amharic ASR model. In our experiments, we transfer acoustic models trained on two different source languages (English and Mandarin) to Amharic using very limited resources. The experimental results show that a significant WER (Word Error Rate) reduction has been achieved by transferring the hidden layers of the trained source languages neural networks. In the best case scenario, the Amharic ASR model adapted from English yields the best WER reduction from 38.72% to 24.50% (an improvement of 14.22% absolute). Adapting the Mandarin model improves the baseline Amharic model with a WER reduction of 10.25% (absolute). Our analysis also reveals that, the speech recognition performance of the adapted acoustic model is highly influenced by the relatedness (in a relative sense) between the source and the target languages than other considered factors (e.g. the quality of source models). Furthermore, other Semitic as well as Afro-Asiatic languages could benefit from the methodology presented in this study.

Keywords: Less-resourced, Semitic languages, Amharic, Transfer learning, weight transfer, Automatic Speech Recognition

1. Introduction

Afro-Asiatic is one of the major language families widely spoken in north and west Africa. Semitic languages belong to Afro-Asiatic. Next to Arabic, Amharic is the second most spoken Semitic language. Moreover, Amharic is an official language of Ethiopia, spoken by over 22 million people, according to Central Statistical Agency of Ethiopia[1]. Amharic has its own unique orthographic representation containing 32 consonants and 7 vowels called **Amharic-Fidel**. The orthographic representation is also shared with Tigrinya, the other Semitic language of Ethiopia (also the main language of Eritrea). Amharic also shares several linguistic features (including morphological structure and vocabulary) with Arabic.

Although there is a large volume of Amharic content available on the web, searching and retrieving them is hard as they only exist in their raw form (not analyzed and indexed well). Therefore, building language specific tools that analyze and index, could potentially enhance the accessibility of Amharic web content. Particularly, automatic speech recognition highly improves the searchability of audio and video content due to its speech transcription support (Mezaris et al., 2010).

Existing Amharic ASR prototypes never seem to be used to perform even other common speech oriented tasks such as language learning (Farzad and Eva, 1998) or solve practical problems by integrating them in other large natural language processing systems such as machine-translation. This is mainly due to the requirement of a fairly large amount of annotated data (e.g., speech transcriptions, language models, lexicons) along with a reasonable degree of quality sufficient to train ASR models.

Compared to other well researched languages for which computational linguistic models have been developed, Amharic is one of the less resourced languages due to a lack of research attention. Even though there are some growing efforts to build general multi-lingual ASR systems and resources (Karafiát et al., 2017; Rosenberg et al., 2017; Das et al., 2016) to support low resourced languages, some languages (including Semitic ones) require exclusive attention due to their unique linguistic nature.

There are also some studies (Abate et al., 2009; Tachbelie et al., 2014; Demeke and Hailemariam, 2012; Melese et al., 2017) on developing language-processing technologies for Amharic, but most of them are done with very limited resources (ELRA-W0074, 2014; Gauthier et al., 2016; HaBiT, 2016). They also do not consider re-using linguistic resources available for other languages. As a result they fail to achieve sufficient quality, especially for commercial use.

Developing good quality speech recognizers typically requires large amounts of transcribed speech and texts. Unfortunately, only small quantities of such data are available for Amharic. They are also limited for specific application domains (not diverse) and formatted to work on specific frameworks. Moreover, preparing data is expensive and time-consuming as it needs to be manually annotated. Therefore, in this study, we explore techniques that enable sharing and adapting existing resources available for other languages.

The most widely used approach to alleviate resource related problems is multilingual model training using pooled data from various languages (Ghoshal et al., 2013; Wang and Zheng, 2015; Karafiát et al., 2017), where under resourced languages get trained together with well-resourced ones. Then, the resulting model could serve to recognize inputs of any of these languages (Wang and Zheng, 2015; Feng and Lee, 2018). While multilingual training potentially improves the recognition performance of the under resourced languages, it demands a huge amount of multilingual re-

[1] https://www.csa.gov.et

sources including a universal (shared) phone set, speech-text corpora, language models and lexicons (Besacier et al., 2014; Wang and Zheng, 2015; Karafiát et al., 2018). In addition, the languages need to be somehow similar (related) to achieve a better outcome. It is often challenging to meet these requirements, especially for those languages that have never been investigated through this approach. Moreover, the problem gets worse when it comes to a language family where most of the member languages are under resourced. Semitic is such an example.

The alternative approach that relaxes these requirements is the transfer learning approach (Huang et al., 2013; Ghahremani et al., 2017; Manohar et al., 2017) (explained in Section 2). Once an acoustic model is trained solely on one of well resourced languages (source languages), the model could be adapted to baseline systems built for less-resourced ones (target languages) through transfer learning. Compared to multilingual training, transfer learning does not only eliminate the requirement for the shared phone set, the source and the target languages do not necessarily need to be related. Also, in terms of computing resources, training multiple languages simultaneously is more costly than training them sequentially.

In this paper, we investigate how well transfer learning is effective for improving the performance (regarding accuracy) of the selected under resourced Semitic language (Amharic) ASR. We aim to adapt pre-trained acoustic models built on two well resourced languages: English and Mandarin. In the speech recognition community, these source languages are considered to be widely accepted as resource rich languages for speech recognition research.

Among other Afro-Asiatic languages, Amharic is strongly related with other many Ethiopian and Eritrean Semitic (e.g., Tigrinya) and non-Semitic Afro-Asiatic (e.g., Afar) languages. Thus, the learning transfer methods achieved in this study could potentially be further transfered to several under resourced Ethiopian and Eritrean languages.

In this paper, we discuss related works in Section 2, transfer leaning in Section 3, the experimental setup in Section 4, the results and discussion in Section 5, the challenges and solutions in Section 6 and, finally, future work and conclusion in Section 7.

2. Related Work

Even though it seems to be difficult to find published articles on transfer learning that are targeted directly at Semitic languages, there are a number of studies (Abate et al., 2009; Yifiru, 2003; Tachbelie et al., 2014; Melese et al., 2017) on the development of ASR for Amharic using conventional methods. Also in (Karafiát et al., 2017; Huang et al., 2013; Rosenberg et al., 2017), some European and other low-resourced languages have been investigated using multilingual transfer learning.

A survey study can be found in (Abate et al., 2009), which summarizes the ASR research attempted for Amharic over the years (2001-2015). According to the survey, speech recognition systems ranging from syllable to sentence level detection, from speaker dependent to speaker independent speech recognition, are built. However, most studies only built proof-of-concept prototypes using quite limited data, similar acoustic modeling techniques i.e. HMM (Hidden Markov Model) (Rabiner, 1989) and tools such as HTK (HMM Tool Kit). State of the art methods such as deep learning (neural methods) do not seem to be investigated yet for Amharic, while the survey was conducted.

Compared to other languages where ASR is being used in various speech technology applications, ASR research for Amharic is very young. There are, of course, a few attempts to integrate an Amharic ASR into different applications, for example, the Microsoft Word application to enable hands-free interactions and support speech commands. Also in (Woldemariam, 2018; Karafiát et al., 2017; Rosenberg et al., 2017) some effort has been made to build a deep neural network (DNN) based ASR for Amharic.

For instance authors in (Woldemariam, 2018) design an ASR-named entity recognition (NER) pipeline that serves as a meta-data extractor in a cross-media framework. The ASR-NER pipeline aims to generate speech transcriptions from audio/video content and tags words in the transcriptions with part-of-speech tags. That potentially helps index Amharic content with those tags and improves their searchability (Chang et al., 2005; Le et al., 2017). However, relatively the recognition quality of the ASR is low and needs to be improved further.

Among other alternative ways to improve the speech recognition accuracy such as increasing training data, improving the quality of language models and pronunciation dictionaries, adapting pre-trained acoustic models available for other languages seems to be more reasonable in terms of resource requirements and time. For example, Jui-Ting et al. in (Huang et al., 2013) experimented with neural net based ASR models transferring for European languages (French, German, Spanish and Italian) and achieved relative WER reductions up to 28%.

There are also some attempts (Manohar et al., 2017; Elmahdy et al., 2013) on adapting cross-lingual acoustic models for Arabic. However, the transfer learning methods used in these studies used to just solve speech recognition errors caused by out-domain-data problems. The authors in (Manohar et al., 2017) apply the transfer learning approach to acoustic models trained on a corpus of multi-dialect Arabic TV broadcast to the YouTube video corpus. In their experiments, all the hidden layers from the source model transferred to the target model and the target model gives an 11.35% absolute improvement over the baseline system. The authors in (Elmahdy et al., 2013) investigate the joint training adaptation approach to improve an acoustic model trained on one of the dialects of Arabic i.e, Qutari.

3. Transfer Learning for Less-Resourced Languages

One way of adapting models trained for one domain/language to another is through the transfer learning method (Wang and Zheng, 2015; Huang et al., 2013; Ghahremani et al., 2017). Parameters learned by a pre-trained deep neural net based model can be transfered to new domains/languages. These parameters are neural net weights estimated and computed during model training.

In natural language processing (NLP), this approach can be applied to transfer knowledge between models trained

on data of different related languages. For example, Greg et al. in (Durrett et al., 2012) applies transfer learning in dependency parsing by using bilingual lexicons of two different languages, acting as source and target. The authors make syntactic analysis of parallel sentences of resource-rich and resource-poor languages, to transfer learned syntactic knowledge between words representing similar concepts. For instance, if there are two words (that mean the same thing) in English and German sentences, the contextual syntactic information of the word in English, could be transferred to the word belonging to German, though not always applicable (effective). In dependency parsing, that potentially used to determine lexical attachment choices during a syntactic tree construction. By using this idea, the authors in (Durrett et al., 2012) reported that significant gains have been achieved for some target languages.

Transfer learning has also been effectively used in speech recognition applications, to adapt acoustic models trained for resource-rich domains (or well-resourced languages) to under-resourced domains (or less-resourced languages).

The main advantages of the adaptation is to tackle resource scarcity and reduce the effort of preparing a huge amount data which is always a challenge in speech recognition research. Moreover, as is evident from some studies (Ghahremani et al., 2017; Yan et al., 2018; Zhuang et al., 2017; Feng and Lee, 2018), the resulting transfered acoustic models perform better as long as the source models perform well and are related with target languages/domains.

Unlike other adaptation methods such as multilingual training, transfer learning does not necessarily require a shared universal phone set across languages. Multilingual training performs multitasking training, that includes merging data from source and target languages, and build a shared acoustic model where each language has its own final (softmax) layer. On the other hand, transfer learning does not necessarily require phone set matching. That practical reason makes it preferable for less-resourced languages, particularly those whose phone set is very unique and hard to share with others.

Compared to other under resourced European languages where multilingual/cross-lingual model adaptation is quite applicable due to their relatedness, Semitic languages seem to have their own unique phone sets along with phonetic representations that are hard to map with other languages. Thus, employing the learning transfer approach seem to be a reasonable choice to serve under resourced Semitic languages.

Knowledge transfer in transfer learning can be achieved by sharing hidden layers of already trained neural net based models. While the input and final layers of pre-trained models are assumed to be language dependent, the hidden layers are regarded as language independent and transferable between languages.

During the learning transfer process, the final layer gets removed from the source models and replaced by the final layer of the model being trained for the corresponding target languages. Also, the input layer gets trained on the data of the target languages. Finally, the whole network is retrained with the shared hidden layers and evaluated on the target language test set (Zhuang et al., 2017; Feng and Lee, 2018).

Generally speaking, transfer learning in speech recognition can be summarized with the four steps: building acoustic models on source languages, removing the final layers from the trained models, transferring hidden layers to target languages and re-train the models with new data.

In practice, however, several challenges may occur during the application of transfer learning, these include mismatching between source and target languages in many ways such as variations in extracted acoustic features. Unless properly handled, these mismatching potentially lead to a high speech recognition error.

4. Experimental Setup

Our experiments cover three different languages: Amharic, English and Chinese Mandarin. While Amharic is intended to be a target language, English and Chinese are source languages.

Kaldi (Povey et al., 2011) has been used as an open speech recognition toolkit for ASR prototypes development and evaluation. It has been configured with the CUDA toolkit to access the GPU card (GeForce GTX 1050 Ti) installed on our machine and train DNN models on the selected source and target languages.

4.1. Datasets

An Amharic corpus consists of read speech collected from 100 different Amharic native speakers of 20 hours for training and 2 hours for testing. Information regarding gender distributions across the speakers is not provided in the paper (Tachbelie et al., 2014) where the corpus with its lexicon is prepared as experimental data. As part of our study, we built different size n-gram language models (n=3 to 7) using the SRILM[2] language modeling toolkit.

An English corpus (Panayotov et al., 2015)] consists of two sets (100 and 360 hours) of read speech (the majority have the US English accent) prepared from audio books, collected from OpenSLR (open speech and language resources)[3]. The test set contains 5.4 hours of speech. We run two different experiments corresponding to the two sets of speech corpus and build acoustic models on each set. In order to assess how the amount of training data affects the result of transfer learning, the first experiment is done with the acoustic model trained on the 100 hours (English-1) set and the second one involves combining the two sets (English-2) together, (we refer English-1 to the 100 hours set and English-2 to the 460 hours set).

A Mandarin corpus (Bu et al., 2017)] contains 178 hours (of which 85% is for training and the remaining is for testing) of speech collected from 400 speakers, provided by Beijing Shell Technology[4] as an open source database. That is the largest Mandarin corpus available for ASR research (Bu et al., 2017) and can be found at OpenSLR[5].

[2]http://www.speech.sri.com/projects/srilm/

[3]http://www.openslr.org/12/

[4]http://www.aishelltech.com/kysjcp

[5]http://www.openslr.org/33/

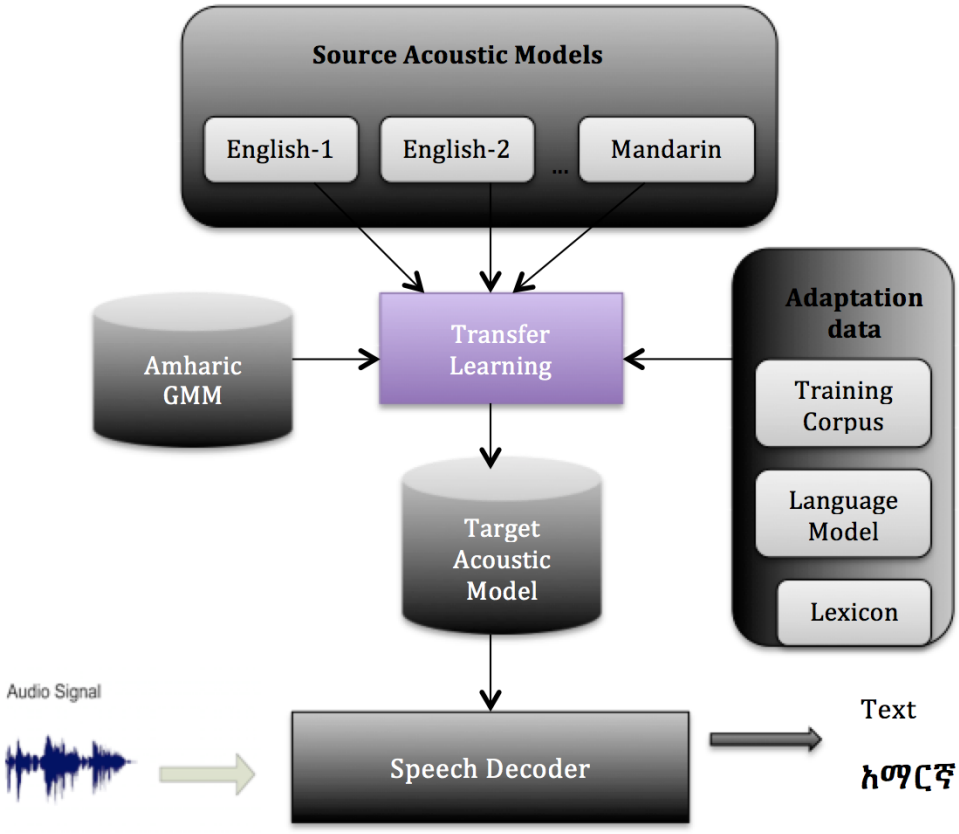

Figure 1: The transfer learning based architecture of the proposed Amharic ASR

4.2. Baseline Systems

A deep neural net baseline system has been built for each language and evaluated on their test sets (the results are summarized in Table 1) after the development of context dependent GMM-HMMs (Gaussian mixture model-hidden Markov model) acoustic models.

The GMM-HMM models are tri-phone based intermediate acoustic models, generated after mono-phone models training. And they are used for the purpose of doing initial alignment (speech with text) for the DNN training.

The DNN acoustic models trained on English and Mandarin, are used for as seed models to be adapted to Amharic. These models make use of a TDNN (time delay deep neural network) architecture (Peddinti et al., 2015) along with the ReLU (rectified linear unit) and 6 hidden layers, each layer has 1026 units. The TDNN architecture is capable of capturing wider context information, and is more efficient than other DNN architectures (e.g. recurrent neural networks).

During training, each frame of the input data is provided with 5 preceding and 5 succeeding frames as contextual information to the network. In order to optimize model parameters (weights and biases), the stochastic gradient descent algorithm is used and run iteratively over the development set. Prior to that, features required to train acoustic models are extracted from the speech corpus of each language. These include MFCC (Mel-Frequency Cepstrum Coefficients) and I-vector (George et al., 2013) features. In addition, speaker independent features are extracted using LDA (Linear Discriminant Analysis) and MLLT (Maximum Likelihood Transform) techniques (Gales, 1998; Gopinath, 1998).

4.3. Transfered Models

Before the actual transfer learning process, we did feature dimension matching between the source languages and the target languages. That is achieved by taking the matrix of the source languages produced at the LDA stage, providing to the target language to re-train their LDA model. We also need to make sure that they have the same splicing settings which determine the context size of concatenated speech segments. For example, while the Chinese corpus uses the splicing options of ”–left-context=5 –right-context=5”, relatively Amharic uses narrow context i.e. ”–left-context=3 –right-context=3”.

Compared to the English model, preparing and adapting from the Mandarin model is quite difficult as it uses very different acoustic features and parameter settings than Amharic and English, due to its tonal nature, (discussed in detail in the challenges and solutions section).

Then we provided the transfer learning algorithm the two required inputs for generating transfered models: pre-trained acoustic models of the sources languages (English and Mandarin), and the adaptation data from the target language (Amharic) along with its GMM-HMM model (as il-

lustrated in Figure 1). The adaptation data includes the speech and text corpus, the lexicon and the language model prepared for Amharic.

The learning algorithm, then takes each pre-trained acoustic model at a time and removes their final (softmax) layer and transfers all the hidden layers to the target Amharic model. Once the transfer has been made, the final layer of the target model gets trained on the adaptation data and added on top of the transferred layers. Also, the weights and biases of the resulting acoustic model is re-computed and fine-tuned with back-propagation. As part of the target model fine-tuning, a smaller learning rate (compared to the learning rate set to the source models) has been used. Among other hyper-parameters (e.g, batch size, number of transferred layers, and so on), lowering the learning rate seems to gives a better result (Ghahremani et al., 2017; Ghoshal et al., 2013). Finally, each version of the final transfered model has been evaluated on the Amharic test set. The results from the transfer learning algorithm have been summarized in Table 1.

5. Results and Discussions

As the experimental results shown in Table 1, the recognition performance of the baseline acoustic model trained for Amharic has been significantly improved through transfer learning. In the best case scenario, adapting from English-2 (with 460 hours), yields improvement over the baseline Amharic ASR with WER decreasing from 38.72% to 24.50% (14.22% absolute). Next, a significant (absolute) improvement is achieved by the model transfered from Engilsh-1 (with 100 hours) to Amharic by 13.66%. The Mandarin model gives a 10.25% absolute reduction over the baseline Amharic model.

Baseline Models	**WER(%)**	
Amharic	38.72	
English-1	8.06	
English-2	**5.75**	
Mandarin	14.65	
Adapted Models		
Source Model	**Target Model**	**WER%**
English-1	Amharic	25.06
English-2		**24.50**
Mandarin		28.47

Table 1: Experimental results from the baseline and the adapted acoustic models

We attempt to analyze the results across three important parameters: the relatedness between the source and the target languages, the quality of the source models, the amount of the data used to train the source languages. We also consider other possible independent factors that potentially influence the performance of the adapted acoustic model and provide analysis on phonetic similarities/ differences of source-target models.

5.1. Impacts and Implications of Source-Target Models Relatedness over Speech Recognition

Basically, in transfer leaning there is a general intuitive assumption that, more or less, natural languages share similar characteristics and are guided by common linguistic principles (Wang and Zheng, 2015). That leads transfer learning to be carried out between two unrelated languages, though more effective when the source and the target languages are somehow similar.

When it comes to this study, assessing how the similarity of the source languages (English and Mandarin) with Amharic impacted the speech recognition performance of the resulting acoustic model is not easy, as there is no direct relationship between them in terms of phonology. While Amharic is one of the most phonetic languages, English and Mandarin are viewed as non-phonetic. That means, in the phonetic languages, a grapheme (alphabetic letters) always has the same sound regardless of its context, whereas in the non-phonetic languages, a single phoneme might have multiple phone realization (variants) depending on its context. Of course, some Amharic speech units have various orthographic representations, but such variations do not affect meanings of words.

Furthermore, there is not sufficient literature that clearly show that how the phonology of such languages associated with Amharic. Relatively speaking, while there are a few investigations (Gashaw, 2017; Yimam, 2000) on phonetic similarity between Amharic and English, there does not seem to exist any studies between Amharic and Mandarin. For instance, Judith et al. in (Judith et al., 2008) show Amharic incorporates several English loan words into its vocabulary, these words are mainly from medical and technology domains. That somehow increases the chance of shared medical or technological terms for being correctly recognized by the adapted acoustic models. In relation to that, however, the size of the Amharic lexicon used in this study is quite small (i.e., 65k), for example, compared to English (i.e., 130k).

Also, the grapheme-to-phoneme (letter-to-sound) rules used in the lexicon do not seem to capture complex syllabification phenomena (e.g., geminnation, presence of the epenthetic vowel) that typically occur in the Amharic phonetics (Hailu and Hailemariam, 2012; Sebsibe et al., 2004; Demeke and Hailemariam, 2012). Obviously, that causes the OOV (out-of-vocabulary) words (new unseen words that do not belong to a lexicon) problem and increases the recognition error during decoding. As observed from the decoding results in Table 1, the performance of the acoustic models trained on such source languages have been impacted by their lexicon size and show differences over Amharic. So, increasing the Amharic lexicon's size could minimize the effect of the OOV words problem and reduce recognition errors. In addition to that, improving the quality of the lexicon using a grapheme-to-phoneme converter that better detects syllable structures of Amharic words might enhance the recognition performance of the Amharic acoustic models.

5.1.1. Phonetic Inventory Overlapping between Source and Target Models

One of the most important aspects of source-target models' relatedness is the similarity between them at the phone level, as the potential underlying reasons for speech recognition errors of the adapted acoustic models might be pretty much related with phonetic mismatching between the target and the source models (Wang and Zheng, 2015; Huang et al., 2013; Ghahremani et al., 2017). During weights (model parameters) transfer, the internal (transferable) DNN layers get trained on source models' phone sets. So, ideally having similar phone sets between source and target languages highly improves the quality of the target models. However, while that is not the case between Amharic and Mandarin, there is partial overlapping between Amharic and English (Gashaw, 2017). Baye in (Yimam, 2000) reveals some similarities between speech units (vowels and consonants) of Amharic and English including their articulation.

Further investigations and understanding of similarities (especially between Amharic and other tonal languages including Mandarin) at the phone level would be interesting as future directions to effectively benefit out of transfer learning.

Looking into the corpus structure used, and domains covered by English and Mandarin, while there are some similar features (e.g. sampling frequency and audio recordings quality) shared between them, the English corpus contains only read speech and the Mandarin corpus mixes both read and telephone speech. On the other hand, the Amharic speech corpus (Tachbelie et al., 2014) contains read speech. That might slightly cause bias towards English. Thus, these facts provide us important clues why adapting the English acoustic model to Amharic is more effective in reducing speech recognition errors.

5.2. Impacts and Implications of the Quality of Source Models over Speech Recognition

The other most important factor is the quality of the source acoustic models. However, considering that the two source models are evaluated on different test sets, it is hard to exactly measure the quality difference between them. Thus, we take the comparisons made between the quality of the acoustic models of English and Mandarin in a relative sense. As shown in Table 1, the baseline systems of English outperform the Mandarin model. As also observed from the WER results of the target models, the Amharic models transferred from English outperform the models transferred from Mandarin. That partially indicates how the quality of the source models affects the quality of the target model.

Probably, that is not always the case because tonal languages like Mandarin can be enhanced by using pitch features and those features have much less influence on non-tonal target languages like Amharic. However, as very significant part of prosodic information (e.g., duration, intonation) of speech, adding the pitch features potentially helps for capturing emotions in speech for both the source and target languages. For instance, effectively detecting such information by acoustic models used, for instance in spoken dialog applications, leads to better decisions (Min and Shrikanth, 2005) during human-machine communications. Therefore, different results might be obtained, if these features are included in both source and target languages.

5.3. Impacts and Implications of the Data Size of Source Models over Speech Recognition

In general, regardless of acoustic models adaptation, the quality of any ASR system is heavily dependent on the quantity of the training data. That is also true in case of transfer learning model adaptation. In our experiment, the model trained on a largest data i.e., English model, has the lowest WER, whereas, the model trained on the smallest data i.e., Amharic, has the highest WER. Also, the target model transfered from the source model trained on the largest dataset yields the best WER. However, the training data size of the Mandarin model is greater than the English-1 by 78 hours, yet the source model trained on English-1 outperforms the Mandarin one by a WER of 6.59%. Also, the transfered model from the Mandarin has slightly higher WER than English-1. This indicates that, the recognition performance of the target models seem to be more sensitive to the quality of the source models than the quantity of the data set where the source models are trained on.

5.4. Impacts and Implications of Other Independent Linguistic Factors over Speech Recognition

It is also worth considering other linguistic factors that are pertinent for understanding the cause of the target model recognition errors. Some of the factors are pretty much inherent to the linguistic and phonetic nature of Amharic, which also apply to other Semitic languages.

Morphologically, Amharic is highly inflectional and complex. That implies, a single Amharic word could appear in many alternative forms conveying various lexical meanings. Like any NLP systems, the Amharic ASR is affected by such morphological complexity. Moreover, as discussed above the Amharic lexicon used in this study is too limited to handle words coming in various derivations. That potentially leads to the OOV problem. To partially address such problem the text corpus containing speech transcriptions has been segmented into morphemes. Also the entries of the lexicon and the language model are made to be morpheme based. Although such approach helps achieve a reasonable performance improvement over word-based ASR, it still gets challenged with OOV words unless supported with a high quality morphology analyzer.

Among other speech sounds in Amharic that could affect the quality of acoustic models, possibly leads to speech recognition errors is the epenthetic vowel (i.e., /ix/) (Sebsibe et al., 2004). While being present in spoken words or utterances, mostly absent in the corresponding training transcriptions causes acoustic confusability. Effective handling of such vowel during acoustic models building takes a bit of research effort, particularly in the context of speech recognition.

6. Challenges and Solutions

In our experiments, compared to English models, adapting from Mandarin seems to be a bit complex and requires more effort due to the presence of extra dimensions (added to

capture the tonal nature of Mandarin) in the trained acoustic network. Originally, the corpus is prepared to have 43 dimensions, that quite deviate from the standard followed to develop ASR for other languages. There are at least two alternative solutions: either adjusting the dimensionality of the adaptation data or reducing the dimensionality of the features by which the network trained on. Relatively the former seems to be difficult as it affects the target language and takes a bit of effort than the later option. We, therefore, took the later option in order to solve the problem and align with the dimension of the adaptation data used by Amharic.
Investigating speech recognition methods, particularly transfer learning is very expensive in many ways. Because most of transfer learning related studies (Karafiát et al., 2017) are based on the proprietary speech corpora mostly purchase from LCD (Linguistic Data Consortium)[6]. Even worse, they released data for some selected languages. For instance, while it is possible to get LDC datasets for other low-resourced languages (e.g., Swahili) with a reasonable price, the separate Amharic datasets are not released yet. The one which is available (by the time this research has been conducted) in LDC is packed with other languages, and to buy the whole pack is really quite expensive.
Therefore, our study is limited to the data available from open source providers. For this reason, in our experiments, relatively well resourced Semitic languages e.g., Arabic, are not considered as source languages. Moreover, that affects the flexibility of our experiments, and experimenting with other variants of the transfered learning approach is quite difficult.
Although transfer learning seems to be a good alternative approach to deal with the problem of resource scarcity, it heavily depends on several pre-conditions that need to be met in advance. Satisfying these requirements, in turn, become challenging in terms of time and cost.
Apart from failing to tackle some mis-match conditions between source and target languages, the lack of deep neural net based computing resources (e.g., GPU) needed for extremely large matrix operations seriously affect the expected results. To meet such challenge and be able to run transfer learning experiments on our server, we took different actions, reducing the size of frames processed at a time, the number of training/decoding jobs and so on. Our experiments have been based on an exclusive use of a single GPU processor with limited memory. That is only able to run one job at a time. That affected the experiments in many ways, for example, limiting the training with certain parameter settings (instead of trying to use possible alternative parameters) and slowing down the training processes, in particularly training larger acoustic models (e.g, the English model with 460 hours takes about a week).

7. Conclusions and Future Work

We conducted transfer learning experiments with selected source and target languages. As a result, we demonstrate that transfer learning could improve the recognition performance of the selected Semitic language. We also attempted to assess the factors affecting the quality (speech recognition performance) of the results obtained from transfer learning. Our assessment partially reveals that, the relatedness (in a relative sense) of the source languages with the target language has high impact than other related factors discussed in Section 4. Due to this reason, the Amharic ASR models transfered from English outperform the model transfered from Mandarin. Also within English source models, the model trained on the larger data set gives better recognition performance.
According to our experimental results, transfer learning seems to be a very effective method as long as the preconditions discussed above are sufficiently met. However, most under-resourced Semitic languages did not take advantage of such recently introduced model adapting methods due to various reasons. We think that this research effort sheds light for investigating transfer learning for other related Semitic languages such as Tigrinya, Arabic, Hebrew and so on. Thus, in the future, it is very interesting to further explore how these languages benefit from the transfer learning approach. Moreover, we consider to investigate other model adaptation methods, in particular multilingual training with additional open source multilingual data and powerful computing resources.
It would also be interesting to evaluate how well the resulting acoustic models perform in various speech based applications such as machine translation, and media analysis frameworks.

[6]https://www.ldc.upenn.edu/

8. Acknowledgments

We acknowledge the financial support from the Kempe foundation, Sweden.

9. Bibliographical References

Abate, S., Tachbelie, M., and Menze, W. (2009). Amharic speech recognition: Past present and future. In *In: Proceedings of the 16th International Conference of Ethiopian Studies*, pages 1391–1401.

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

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Chang, S.-F., Manmatha, R., and Chua, T.-S. (2005). Combining text and audio-visible features in video indexing. In *Acoustics, Speech, and Signal Processing*, pages 1005–1008.

Das, A., Jyothi, P., and Hasegawa-Johnson, M. (2016). Automatic speech recognition using probabilistic transcriptions in swahili, amharic, and dinka. In *INTERSPEECH*, pages 3524–3528.

Demeke, Y. and Hailemariam, S. (2012). Duration modeling of phonemes for amharic text to speech system. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES)*, pages 1–7.

Durrett, G., Pauls, A., and Klein, D. (2012). Syntactic transfer using a bilingual lexicon. In *EMNLP-CoNLL*.

Elmahdy, M., Hasegawa-Johnson, M., and Mustafawi, E. (2013). A transfer learning approach for under-resourced arabic dialects speech recognition. In *Proceedings of the 6th Language and Technology Conference*, page 290â293.

Farzad, E. and Eva, K. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm. *Language Learning & Technology*, 2(1):45–60.

Feng, S. and Lee, T. (2018). Improving cross-lingual knowledge transferability using multilingual tdnn-blstm with language-dependent pre-final layer. In *Interspeech*, pages 2439–2443.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Science and Language*, 12:75–98.

Gashaw, A. (2017). Rhythm in ethiopian english: Implications for the teaching of english prosody. *International Journal of Education and Literacy Studies*, 5(1):13–19.

George, S., Hagen, S., David, N., and Michael, P. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 55–59.

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

Ghoshal, A., Swietojanski, P., and Renals, S. (2013). Multilingual training of deep neural networks. *2013 IEEE International Conference on Acoustics Speech and Signal Processing*, pages 7319–7323.

Gopinath, R. (1998). Maximum likelihood modeling with gaussian distributions for classification. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, pages 661–664.

Hailu, N. and Hailemariam, S. (2012). Modeling improved syllabification algorithm for amharic. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 16–21.

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

Judith, R., Rotem, K., and MyiLibrary. (2008). *Globally speaking : motives for adopting English vocabulary in other languages*. Clevedon UK ; Buffalo [N.Y.] : Multilingual Matters, New York.

Karafiát, M., Baskar, M. K., Matejka, P., Veselý, K., Grézl, F., Burget, L., and Cernocký, J. (2017). 2016 but babel system: Multilingual blstm acoustic model with i-vector based adaptation. In *The Proceedings of INTERSPEECH 2017*, pages 719–723.

Karafiát, M., Baskar, M. K., Veselý, K., Grézl, F., Burget, L., and ernocký, J. (2018). Analysis of multilingual blstm acoustic model on low and high resource languages. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5789–5793.

Le, N., Bredin, H., Sargent, G., India, M., Lopez-Otero, P., Barras, C., Guinaudeau, C., Gravier, G., da Fonseca, G. B., Freire, I. L., do Patrocínio, Z. K. G., Guimarães, S. J. F., Martí, G., Morros, J. R., Hernando, J., Fernández, L. D., García-Mateo, C., Meignier, S., and Odobez, J.-M. (2017). Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proceedings of International Workshop on Content-Based Multimedia Retrieval*, pages 1–6.

Manohar, V., Povey, D., and Khudanpur, S. (2017). Jhu kaldi system for arabic mgb-3 asr challenge using diarization audio-transcript alignment and transfer learning. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 346–352.

Melese, M., Besacier, L., and Meshesha, M. (2017). Amharic-english speech translation in tourism domain. In *SCNLP@EMNLP 2017*.

Mezaris, V., Gidaros, S., Papadopoulos, G. T., Kasper, W., Steffen, J., Ordelman, R., Huijbregts, M., de Jong, F., Kompatsiaris, Y., and Strintzis, M. G. (2010). A system for the semantic multimodal analysis of news audio-visual content. *EURASIP Journal on Advances in Signal Processing*, 2010:1–16.

Min, L. and Shrikanth, N. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *The Proceedings of INTERSPEECH 2015*, pages 3214–3218.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. K., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rosenberg, A., Audhkhasi, K., Sethy, A., Ramabhadran, B., and Picheny, M. (2017). End-to-end speech recognition and keyword search on low-resource languages. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284.

Sebsibe, H., Prahallad, K., Alan, B., Rohit, K., and Rajeev, S. (2004). Unit selection voice for amharic using festvox. In *Fifth ISCA Workshop on Speech Synthesis*, pages 103–107.

Tachbelie, M., Abate, S., and Besacier, L. (2014). Using different acoustic lexical and language modeling units for asr of an under-resourced language - amharic. *Speech Communication*, 56:181–194.

Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. *2015 Asia-Pacific*

Signal and Information Processing Association Annual Summit and Conference (APSIPA), pages 1225–1237.

Woldemariam, Y. (2018). *Natural Language Processing in Cross-Media Analysis*. Licentiate thesis, Faculty of Science and Technology, Umeå University, Jun.

Yan, J., Yu, H., and Li, G. (2018). Tibetan acoustic model research based on tdnn. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 601–604.

Yifiru, M. (2003). Automatic amharic speech recognition system to command and control computers. Master's thesis, School of Information Studies for Africa, Addis Ababa University.

Yimam, B. (2000). *Amharic Grammar*. Eleni Publishing Ltd, Addis Ababa.

Zhuang, X., Ghoshal, A., Rosti, A.-V., Paulik, M., and Liu, D. (2017). Improving dnn bluetooth narrowband acoustic models by cross-bandwidth and cross-lingual initialization. In *INTERSPEECH*, pages 2148–2152.

10. Language Resource References

Hui Bu and Jiayu Du and Xingyu Na and Bengu Wu and Hao Zheng. (2017). *Aishell ASR corpus*. provided by Beijing Shell Technology and distributed via OpenSLR, ISLRN SLR33.

ELRA-W0074. (2014). *Amharic-English bilingual corpus, distributed via ELRA,1.0*. distributed via ELRA,1.0, 1.0, ISLRN 590-255-335-719-0.

Elodie Gauthier and Laurent Besacier and Sylvie Voisin and Michael Melese and Uriel Pascal Elingui. (2016). *Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof*. European Language Resources Association (ELRA), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

HaBiT. (2016). *Harvesting big text data for under-resourced languages*. distributed via Natural Language Processing Centre, Faculty of Informatics, Masaryk University.

Vassil Panayotov and Guoguo Chen and Daniel Povey and Sanjeev Khudanpur. (2015). *LibriSpeech ASR corpus*. distributed via OpenSLR, ISLRN SLR12.