Scientific
Research
Publishing

Journal of Signal and Information Processing, 2020, 11, 1-21
https://www.scirp.org/journal/jsip
ISSN Online: 2159-4481
ISSN Print: 2159-4465

Investigation of Automatic Speech Recognition Systems via the Multilingual Deep Neural Network Modeling Methods for a Very Low-Resource Language, Chaha

Tessfu Geteye Fantaye[1], Junqing Yu[1,2*], Tulu Tilahun Hailu[1]

[1]School of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan, China
[2]Center of Network & Computation, Huazhong University of Science & Technology, Wuhan, China
Email: tessfug@hust.edu.cn, *yjqing@mail.hust.edu.cn, tutilacs@yahoo.com

How to cite this paper: Fantaye, T.G., Yu, J. and Hailu, T.T. (2020) Investigation of Automatic Speech Recognition Systems via the Multilingual Deep Neural Network Modeling Methods for a Very Low-Resource Language, Chaha. *Journal of Signal and Information Processing*, 11, 1-21.
https://doi.org/10.4236/jsip.2020.111001

Received: December 12, 2019
Accepted: January 6, 2020
Published: January 9, 2020

Abstract

Automatic speech recognition (ASR) is vital for very low-resource languages for mitigating the extinction trouble. Chaha is one of the low-resource languages, which suffers from the problem of resource insufficiency and some of its phonological, morphological, and orthographic features challenge the development and initiatives in the area of ASR. By considering these challenges, this study is the first endeavor, which analyzed the characteristics of the language, prepared speech corpus, and developed different ASR systems. A small 3-hour read speech corpus was prepared and transcribed. Different basic and rounded phone unit-based speech recognizers were explored using multilingual deep neural network (DNN) modeling methods. The experimental results demonstrated that all the basic phone and rounded phone unit-based multilingual models outperformed the corresponding unilingual models with the relative performance improvements of 5.47% to 19.87% and 5.74% to 16.77%, respectively. The rounded phone unit-based multilingual models outperformed the equivalent basic phone unit-based models with relative performance improvements of 0.95% to 4.98%. Overall, we discovered that multilingual DNN modeling methods are profoundly effective to develop Chaha speech recognizers. Both the basic and rounded phone acoustic units are convenient to build Chaha ASR system. However, the rounded phone unit-based models are superior in performance and faster in recognition speed over the corresponding basic phone unit-based models. Hence, the rounded phone units are the most suitable acoustic units to develop Chaha ASR systems.

*Corresponding author.

Keywords

Automatic Speech Recognition, Multilingual DNN Modeling Methods, Basic Phone Acoustic Units, Rounded Phone Acoustic Units, Chaha

1. Introduction

Human language technologies (HLTs) are important for the low-resource languages, to revitalize and document them for preventing the challenge of extinction, and to raise the interest and make the language attractive again for their native speakers [1]. ASR is one of the HLTs that is developed for such languages using small training corpora, which are often prepared by researchers. Thus, the performance of speech recognizers of low-resource languages is worse than that of speech recognizers of technologically favored languages. Besides, due to the shortage of sufficient training corpora, the DNN models suffer from overfitting problem when developing speech recognizers for low-resource languages. The scarcity of the training dataset and overfitting challenges of DNN models are mitigated by either increasing the size of the training datasets or developing optimal DNN models using various model regularization techniques such as dropout, l2-normalization, activation functions, layer normalization, and batch normalization.

The model regularization techniques can reduce the overfitting problem to some extent, but to overcome the above problems substantially and to develop reliable ASR systems for the low-resource languages, it is better to increase the size of the training datasets. The size of the training datasets can be increased by preparing a new training corpus, borrowing from high-resource languages, and generating synthetic datasets via various audio data augmentation techniques. The first approach is expensive because it takes considerable time, human, and financial resources and it is challenging to obtain electronically available text for the very much low-resource languages. Thus, it is better to use the second and the third methods, namely, borrow training datasets from the high-resource languages and use the synthetic dataset by generating via data augmentation techniques. Using these methods, different multilingual acoustic modeling paradigms were investigated in the previous works. Phone sharing [2], multitask learning [3] [4] [5], and weight transfer [4] [5] were utilized to develop reliable ASR systems for low-resource languages.

Chaha is one of the low-resource languages, which has limited presence on the web and suffers from lack of language-specific electronic-resources, namely, text corpus, speech corpus, lexical dictionary, and language model, which are used for developing ASR systems. As a result, it is a very low-resource language. Moreover, some of the phonological, morphological, and orthographic features of Chaha challenge the development of ASR system. Due to these problems, there is no study conducted on HLTs in general and ASR system in particular for

Chaha language until now.

This study investigates the development of different speech recognition systems using various multilingual DNN acoustic modeling techniques for the Chaha language, and offers the following contributions:

- Analyzing the characteristics of the Chaha language that favors and challenges the development of the speech recognition systems.
- Developing language resources, namely, text corpus, speech corpus, lexical dictionary, and language model for Chaha language.
- Developing the basic[1] phone and rounded[2] phone unit-based GMM-HMM and unilingual DNN-HMM models for Chaha language.
- Investigating different phone and rounded phone unit-based speech recognizers using various multilingual DNN acoustic modeling paradigms and comparing the recognizers in terms of performance and recognition speed for the Chaha language.
- Comparing and suggesting the best acoustic modeling units to develop speech recognition system for the Chaha language.

The remainder of this paper is organized as follows. The review of related works is presented in Section 2. A description of the Chaha language is given in Section 3. Section 4 describes the preparation of corpora. The experiments, results, and discussion of this work are discussed in Section 5. Section 6 explains the conclusions and future directions of this work.

2. Related Works

Multilingual DNN acoustic modeling paradigms are helpful to share and transfer DNN hidden layers among or between multiple languages for improving the performance of the individual languages. These paradigms are effective to reduce overfitting problem of DNN-based speech recognition systems for low-resource languages. The widely used multilingual DNN acoustic modeling paradigms in speech recognition of low-resource speech recognition systems include phone sharing, multi-task learning, and weight transfer. In phone sharing modeling paradigm, the phones of various languages are either merged with a language identifier prefix or combined with the universal phones of all the languages based on data-driven or International phonetic alphabet (IPA) approaches to create the multilingual phone sets, and then train the model using the mixed multilingual datasets from all languages. For instance, Vu *et al.* [2] have trained two phone sharing multilingual DNN models for the ten languages from Global phone database in the low resource scenarios. The first is merged phone sets based phone sharing, which is created by simply concatenating all involved monolingual phone sets with a language identification prefix to ensure that all the phones are distinct between languages. The second is a universal

[1] Basic phone units contain only basic phones, where the rounded phones are maps to the corresponding basic phones.

[2] Rounded phone units contain all the basic phones and the rounded vowels, where rounded phones map to the basic phones and rounded vowels to consider their roundedness.

phone set based phone sharing, which merges all the monolingual phones that share the same symbol in the IPA table. Using both paradigms, they have obtained superior performances over the corresponding unilingual DNN models.

Multitask learning is helpful to transfer knowledge between or among languages if the languages are phonetically related with each other and share some internal representation by jointly learning together. In this multilingual paradigm, the hidden or initial layers of the network are shared across all languages and each language has a specific output layer, as shown in **Figure 1**. On the other hand, in the weight transfer modeling paradigm, the hidden layers of the source DNN model train using the unilingual or multilingual datasets, and then remove the output layer and replace it with a new target language output layer with dimension equal to the number of senones. Then, train only the added output layer or retrain all the model-hidden layers using small training dataset of the target language, as shown in **Figure 2**. For example, Gales *et al.* [6] have examined the use of shared hidden layer multilingual DNN-HMM models for the low-resource languages from IARPA Babel project. Huang *et al.* [5] have studied the multi-task learning DNN architecture and weight transfer schemes, and attained better performance over the unilingual DNN models. Lin *et al.* [7] have also used these two multilingual DNN models to develop speech recognizers

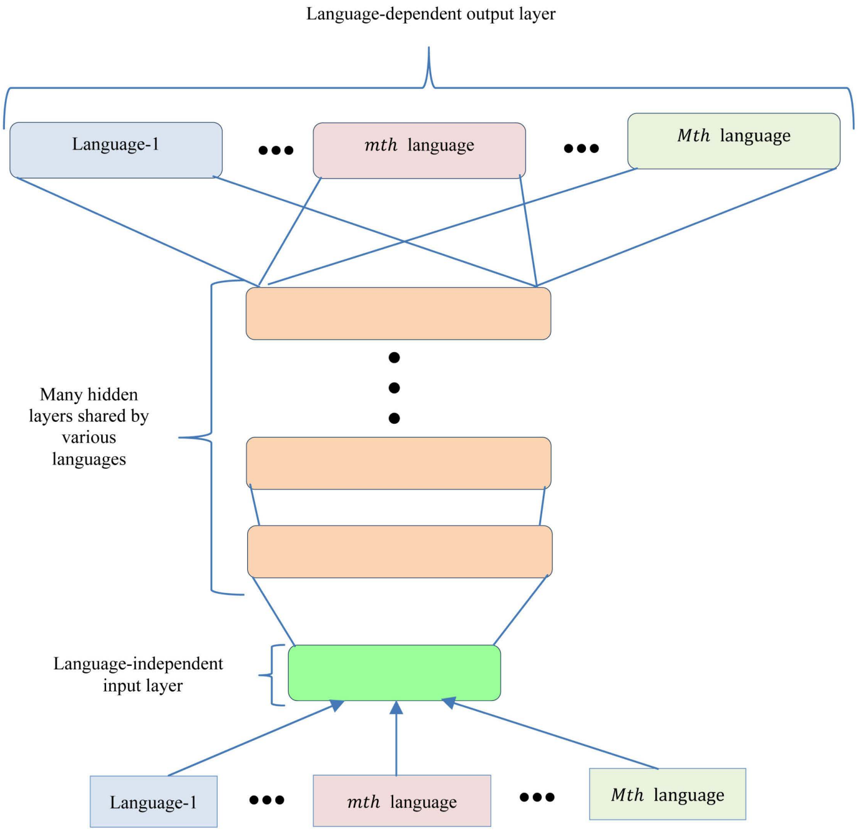
Language-dependent output layer
Language-1
•••
mth language
•••
Mth language
Many hidden layers shared by various languages
Language-independent input layer
Language-1
•••
mth language
•••
Mth language

Figure 1. Multitask or shard hidden layer multilingual DNN paradigm with M languages. A number of hidden layers are shared for multiple M languages and trained via multilingual datasets, while the output layer is specific to each language.

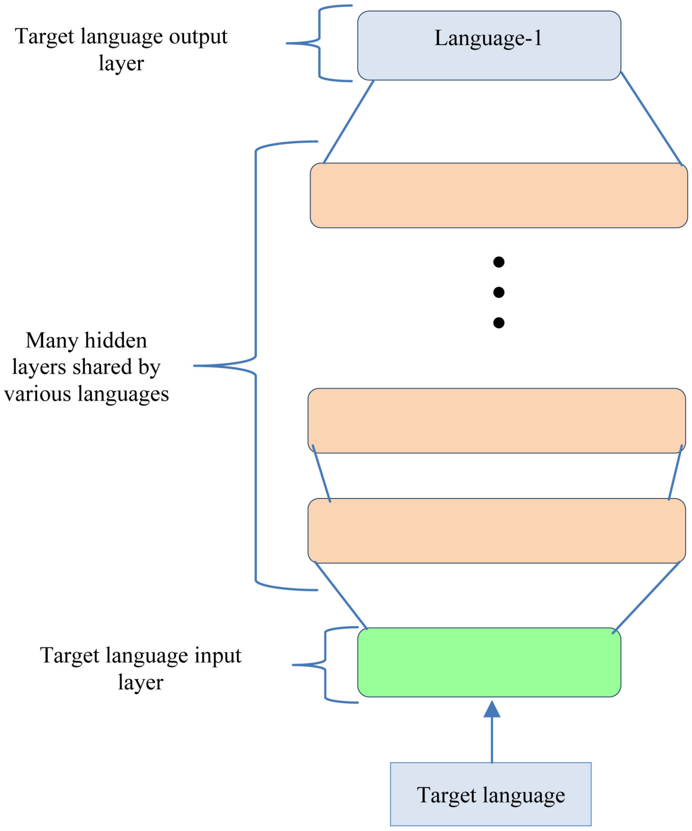
Target language output layer
Language-1
Many hidden layers shared by various languages
Target language input layer
Target language

Figure 2. Weight-transfer DNN paradigm. The hidden layers are borrowed from the multitask or shared hidden layer multilingual DNN, while the output layer requires to be trained with dataset from the target language.

for the low-resource Taiwanese Mandarin language, and obtained better performances using both models, and the multitask learning model outperforms the corresponding weight transfer model. Similarly, Ghahremani *et al.* [4] have compared the multitask learning and weight transfer models using lattice free maximum mutual information (LF-MMI) objective function, and obtained superior performance using the multitask learning model over the weight transfer model. Moreover, Miao and Metze [8] have combined and trained the dropout model regularizer with multitask learning DNN model for the very low-resource language settings, and acquired significant performance improvements.

The performance of the multilingual modeling paradigms is profoundly affected by the size of the training datasets and the relatedness of the languages. Hence, training related target and source languages together produce better performance than training unrelated target and source languages. For example, the works presented in [2] [4] [5], and [8] are trained related target and source languages, and obtained superior performances over the works presented in [2] [5] [7], which train unrelated target and source languages. The target and source languages are considered related languages when they are phonetically related to each other. Commonly, the languages that are found within the same language family are phonetically related languages. For example, Chaha and Amharic are members of the Semitic language family. Hence, these languages are phonetically related to each other.

Different researchers have investigated ASR system for Amharic language [9]. For example, Abate *et al.* [10] have analyzed the language specific and resource-related challenges for developing ASR system for Amharic language. Tachbelie *et al.* [11] have examined syllable and hybrid acoustic modeling units based speech recognizers for Amharic. Tachbelie *et al.* [12] have also analyzed the various acoustic, language, and lexical modeling units to develop Amharic ASR system. However, HLTs in general and ASR system in particular have not been investigated for the Chaha language. Thus, this study is a first attempt to investigate Chaha speech recognition systems using multilingual DNN modeling paradigms by borrowing the training datasets from a phonetically related language, Amharic.

3. The Chaha Language

Chaha is one of the major dialects of the west Gurage language. It belongs to the Semitic language family of which the other members are Arabic, Geez, Amharic, Tigriyna, Argobba, Harari, and Gaft [13]. Chaha is spoken in the Gurage Zone that is located in the southern part of Ethiopia. Gurage settlers in different Ethiopia cities such as Addis Ababa, Dire Dawa, and Hawssa also speak it. Based on a 2007 census, Chaha has around half a million speakers as the first language. This figure does not include a large number of Chaha speakers who live outside the Gurage zone. The linguistic features of Chaha are more studied than the other dialects of the west Gurage language by local and foreigner linguistics [13] [14] [15] [16] [17].

However, Chaha is a developing language. This is because Chaha is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread [18], Chaha is not used as a medium for lesson delivering or as a program in education, namely, primary and secondary schools and higher institutions, and has less documentation and development products. For instance, it has very few books. As of the time of writing this paper only four fictions, one bible, one poem, and one proverb publications are available in the language. Moreover, there are no revitalization efforts and language development agencies for the language. Hence, Chaha needs a particular attention of linguistics and HLTs developers, to make the language easily accessible and usable by the speakers. This section discusses the phonological, morphological, and orthographic characteristics of the language.

3.1. Chaha Phonology

Chaha has 47 speech sound units, which are 38 consonants and nine vowels [13] [14] [15]. The consonants are classified into stops, affricatives, fricatives, spirants, and sonorants based on the manner of articulation, as listed in **Table 1**. The phonetic transcription of the consonants b, p, f, m, w, g, k, d, t, z, s, h, l, n, r,

Table 1. Chaha consonants (adapted from [13] [14]).

Manner of articulation		Place of articulation							
		Labial					Velar		
		Plain	Rounded	Alveolar	Alveo-palatal	Palatal	Plain	Rounded	Laryngeal
Stops	Voiceless ejectives			t’		q’	q	q^w	
	Voiceless	p	p^w	t		k’	k	k^w	
	Voiced	b	b^w	d		g’	g	g^w	
Affricates	Voiceless ejectives				č’				
	Voiceless				č				
	Voiced				ğ				h
Fricatives	Voiceless	f	f^w	s	š				
	Voiced			z	ž				
Spirants						x’	x	x^w	
Sonorants	Nasal	m	m^w	n		ɲ			
	Approximant	*β*	w	r,l		y			

and y corresponds to that of Amharic and English consonants. The pronunciation of consonants t’, č, k’, š, ž, c, ğ, x, β, ɲ, q^w, k^w, and g^w correspond to the equivalent Amharic consonants t’, č’, q, š, ž, č, ğ, h, v, ň, q^w, k^w, and g^w, respectively. The consonant speech sounds, q’, k’, g’, x’, x^w, p^w, b^w, f^w, and m^w are peculiar to Chaha, and do not have corresponding sounds in Amharic and English languages. The sound units q’, k’, g’, and x’ are palatalized counterparts from amongst the consonants q, k, g, and x, while q^w, k^w, g^w, x^w, p^w, b^w, f^w, and m^w are the labialized form of the consonants q, k, g, x, p, b, f, and m, respectively [13] [14]. The only laryngeal sound that exists in Chaha is h, which is used to call a few Amharic loan words such as haymanot “belief”, har “silk”. Chaha is a non-geminating language, in which whenever an originally voiced consonant expects to geminate, it becomes voiceless. For example, the sound b becomes p. However, occasionally one encounters occurrences with a geminated radical, as in ənnəm “all” for the loan words from the Amharic language. Hence, Chaha has only a few consonants that can geminate, namely, m, n, t, and k [13] [15].

The seven basic vowels, namely, ä, u, i, a, e, ə, o, and the two low-mid front (*ɛ*) and back (ɔ) vowels, form the nine phonetic vowel inventory of Chaha, as presented in **Table 2**. The vowel ä alternates with ə vowel of Amharic such as in Chaha äxər meaning “cereal”, while in Amharic it sounds as əhəl. Similarly, the vowels u, i, a, e, and o have the same phonetic transcription with the corresponding Amharic vowels. The vowel ə corresponds to the ɨ vowel of Amharic language. The open ä (ɛ) and open o (ɔ) vowels are distinctive for Chaha, and are minimal pairs with ä and o vowels, respectively.

Most of the Chaha consonants are basic consonant phones, but eight consonant phones (21.1% of the total consonants) are rounded consonants. Thus, the rounded nature of these consonants should be considered during development

Table 2. Chaha vowels (adapted from [13] [14]).

	Front	Central	Back
High	i	ə	u
High-mid	e		o
Low-mid	ɛ	ä	ɔ
Low		a	

of Chaha speech recognition systems. Moreover, there are consonants which have similar pronunciation with different symbols namely, x and h. Nevertheless, the speech recognition system considers a unique pronunciation of orthographic symbols. Hence, one of the two phones should be used when developing Chaha speech recognition system.

3.2. Chaha Morphology

Chaha exhibits a root-pattern, inflectional, and derivational morphological phenomena like other Semitic languages such as Arabic and Amharic [17]. Moreover, Chaha has unique properties, namely, labialization, palatalization, devoicing, and sonorant alternations. Hence, Chaha is a morphologically rich language, and its morphological richness challenges the performances of speech recognition systems.

3.3. Chaha Writing System

Chaha is written using Geez script. However, Chaha represents the palatalized consonants, which are not found in Geez script by introducing the modified characters to the script, such as using wedges on the tops. Chaha script is syllabic, where each symbol represents a consonant combined with a vowel except the sixth-order consonant, which is sometimes realized as a consonant without a vowel, and at other times a consonant with a vowel [13]. Each symbol in Chaha writing system represents a consonant-vowel (CV) syllable, and there are 264 distinct letters [13] [14]: 224 letters—32 core symbols with seven orders, 20 letters—four rounded velars with five orders, and 20 letters—four rounded labials with five orders. In the 32 core letters, consonants are concatenated with every seven basic vowels to obtain a total 264 CV syllables. Similarly, four plain velar letters, namely, q, k, g, and x, and four plain labial letters, namely, p, b, f, and m are combined with five rounded vowels, namely, ^{w}ä, ^{w}i, ^{w}a, ^{w}e, and ^{w}ə, to obtain 20 rounded velars and 20 rounded labial letters, respectively. In addition, Chaha has various syllable structures with a form of C(C)V(C)(C), and the possible syllable types are vowel (V), CV, consonant-vowel-consonant (CVC), vowel-consonant (VC), consonant-vowel-consonant-consonant (CVCC), vowel-consonant-consonant (VCC), consonant-consonant-vowel (CCV), and consonant-consonant-vowel-consonant-consonant (CCVCC) [15]. The CV and CVC syllables are basic, and are called light and heavy syllables, respectively, while the CVCC syllables are the super heavy syllables. CV syllable is the dominantly available syllable type in the language [14] [16].

The phonetic and syllabic features of the Chaha writing system favor the development of ASR systems. For example, it is easy to develop lexical dictionary using a grapheme-based approach. However, Chaha writing system does not show gemination and devoicing of consonants, and pronunciation of an epenthetic vowel ə and open vowels, namely, *ɛ* and ɔ. These characteristics of Chaha writing system are analogous to the vowels of Arabic and Hebrew, and the geminated consonants and epenthetic vowel of Amharic, which are not indicated in writing system. Moreover, Chaha has syllables that have the same pronunciation with different orthographic symbols. Overall, the above features of the language challenge the development of ASR systems.

4. Preparation of Corpora

In this section, the text corpus, speech corpus, lexical dictionaries, and synthetic speech corpora, which we have used in our study, and the process followed to prepare them are discussed.

Chaha does not have a readily available text corpus. Besides, it has limited presence on the web, and has limited hardcopy books. Thus, we have collected small set of texts from bible, web, and hardcopy books such as fiction, poem, and proverbs. Then, the texts of bible, web, and books are merged, and applied text cleaning tasks like correcting spelling and grammar errors, expanding abbreviations, removing foreign words, textually transcribing numbers, and separating concatenated words. As a result, we obtained 14,595 sentences (200,944 tokens and 38,182 word types) as text corpus, which is used to generate lexical dictionaries and to train language models.

Moreover, the phone-level Unicode versions of the text corpus and transcribed speech text are used. The transliteration[3] of the text corpus and the transcribed speech text from their syllable-level Unicode versions into the corresponding phone-level Unicode versions is conducted as follows: All the syllables except the 20 rounded velars and 20 rounded labials syllables are transliterated in terms of CV pattern. For instance, the word በና/bäna/, which means "eat", is transliterated as ብኧንኣ/bäna/, where syllable በ/bä/ is transliterated as the combination of the sixth-order phone, namely, ብ/b/, with the first vowel, namely, ኧ/ä/, to the transliterated form of ብኧ/bä/, and syllable ና/na/ is transliterated as the combination of sixth-order phone ን/n/ with the 4th vowel, namely, ኣ/a/, to the transliterated form of ንኣ/na/. However, the rounded velar and labialized syllables are combinations of two or three CV syllables. Thus, according to [13] [14], these syllables can be transliterated as the concatenations of sixth-order phones with rounded vowels. For example, ኳ/k^{w}a/ is a rounded velar syllable and is transliterated as a combination of sixth-order phone ክ/k/ with rounded vowel ውኣ/^{w}a/ to the corresponding phone transliteration of ክውኣ /k^{w}a/.

Like the text corpus, Chaha does not have publicly available speech corpus for

[3] Transliteration refers to the conversion of a syllable-level Unicode version to the corresponding phone-level Unicode version.

examining speech recognition tasks. Hence, we have prepared the speech corpus by selecting 2000 relatively phonetically balanced sentences from the obtained text corpus. A speech corpus of 3-hour is recorded in an office environment using a Philips voice recorder (VTR5100) from 15 native speakers (10 male and 5 female), who read a total of 2000 sentences. Of the 3-hour speech corpus, 2.67-hour (1778 sentences), is collected from 10 native speakers (7 male and 3 female) who read 178 sentences each. This corpus is utilized as a training dataset. To avoid the overlapping between the training and testing datasets with respect to speakers and sentences, a 0.33-hour (222 sentences) corpus is collected from a separate 5 native speakers (3 male and 2 female) who read 45 sentences each, and this corpus is ten percent of the total 3-hour corpus and is used as a testing dataset. However, compared to other speech corpora that contain tens and above hours of speech data for training, clearly this corpus is very much small, and hence, the models will suffer from lack of training data. The distribution of phonemes within 2.67-hour training dataset is shown in **Figure 3**.

There are no available lexical dictionaries for Chaha language. Hence, we have prepared two basic phone-based and two rounded phone-based lexicons via a grapheme-based approach [19]. The two basic phone-based lexicons contain 36 basic phones: 29 basic consonants and 7 basic vowels, and 32 basic phones: 25 basic consonants and 7 basic vowels, respectively. In the first lexicon, the four palatalized phones are used directly, while in the second lexicon, these phones are mapped into the corresponding basic phones. These lexicons are prepared by a simple transcription of words as separated phones. The two rounded phone-based lexicons consist of 44 phones: 29 basic consonants, 7 basic vowels, 5 rounded vowels and 3 palatial vowels, and 41 phones: 29 basic phones, 7 basic vowels and 5 rounded vowels, respectively. The first lexicon includes additional palatal vowels, while the second lexicon uses the palatal phones directly. These

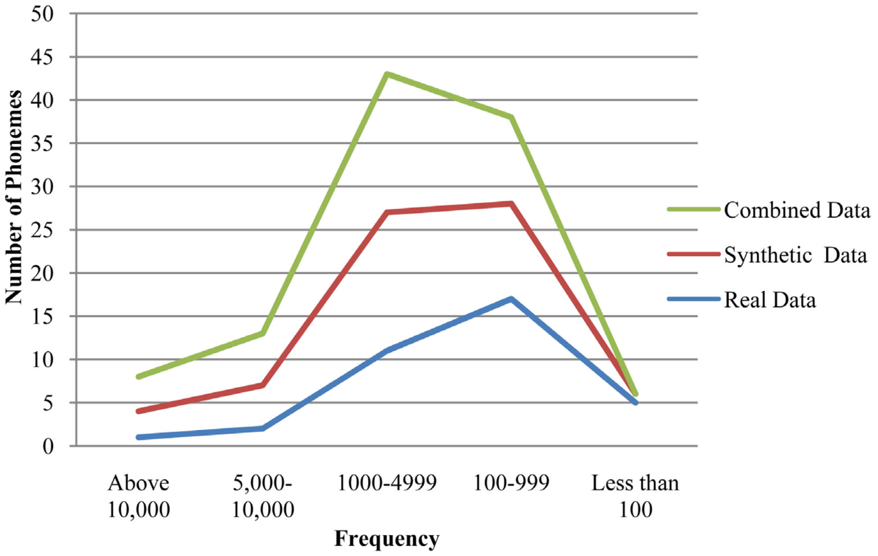

Figure 3. Distribution of Chaha phonemes within real and synthetic training speech corpora.

lexicons are prepared by following the same procedure as basic lexicons except for the rounded phones. These phones transcribe in the lexicon as a separated basic phones and rounded vowels. All lexicons are derived from the text corpus by selecting the most frequent words and are of size 4 k. These lexicons are not considered the language properties, namely, consonant gemination and devoicing, insertion of epenthetic vowel, pronunciation of open vowels, and elision of vowels or insertion of semivowels, because the lexicons are prepared by nonlinguistic experts with the help of the writing system of the language. However, an epenthetic vowel is inserted for all available sixth-order consonant phones in the training corpora and lexical dictionaries.

Moreover, we have used Amharic as a resource provider language. It has 26-hour training speech corpus (from [20] and own), which contains a total of 13,549 sentences that are collected from 125 native speakers. Alternatively, to increase the size of training datasets, the synthetic training datasets are generated using the speed perturbed audio data augmentation approach [21] by modifying the speed of speech signal to 90% and 110% of the initial rate for both languages. **Figure 3** shows the distribution of Chaha phonemes in the synthetic training dataset and **Table 3** lists the summary of the total training datasets used to train the Chaha ASR systems.

5. Experiments

5.1. Experimental Setups

All the GMM-HMM and DNN-HMM models are developed using the state-of-the-art speech recognition toolkit, Kaldi [22]. For GMM-HMM models, speaker adaptive training (SAT) technique based 40-dimensional features are extracted with feature-space maximum likelihood linear regression (fMLLR) method. Various Bakis HMM topology triphone models are built for all basic and rounded phone acoustic modeling units. Moreover, word based backed-off and interpolated trigram language models are built using SRI language modeling (SRILM) toolkit [23]. These language models are smoothed using the modified Kneser-Ney smoothing algorithm, and are applied to train all the basic and rounded phone acoustic units.

For DNN-HMM models, we used a chain model that trains with LF-MMI criterion without the need for frame level cross-entropy pretraining [24]. This model uses a one-state HMM topology for each context-dependent phone, and the phonetic-context decision tree obtains using one-state HMM topology and reduced frame rate after converting the alignments from the GMM-HMM model.

Table 3. Training speech datasets (hours).

Language	Real dataset	Synthetic dataset	Total dataset
Chaha	2.67	5.34	8.01
Amharic	26	52	78

We used the time delay neural network (TDNN) models. We used 40-dimensional high resolution MFCC features and 50-dimensional i-vector speaker adaption features as input features for training the unilingual TDNNs models, and 40-dimensional high resolution MFCC features and 100-dimensional i-vector speaker adaption features as input features for training multilingual TDNNs models. For both models, a left context width of 16 frames and a right context width of 10 frames are used to combine the frames. The ReLU nonlinearity is used for each hidden layers with a batch normalization and dropout model regularization techniques for monitoring the model complexity and overfitting problem. A dropout scheduling with values of 0, 0.2, 0.3, and 0 is used. The L2, xent, and leaky regularization techniques are also applied with values of 0.05, 0.1, and 0.1, respectively. The variable mini-batches, namely, 128, 64, and 32 frames are used for weight updating during training. The training process is accelerated via Nvidia GeForce GTX 1050 GPU on a single machine. A weighted finite state transducer is used for decoding.

In addition to the above universal parameters, the unilingual and weight transfer TDNN models are used output dimensions of 656 and 696 senones for basic and rounded phone units. These models also used similar initial and final learning rate values of 0.004 and 0.0006, respectively. However, the unilingual TDNN models used ten training epochs while weight transfer models applied two epochs. Conversely, phone sharing and multitask learning multilingual TDNN models used the same initial and final learning rate values of 0.0004 and 0.0001, respectively. These models also used six training epochs for both basic and rounded phone units. The output layer dimensions of phone sharing models are 1752 and 1848 for basic and rounded phones, respectively, while the output layer dimensions of the multitask learning models are similar to the unilingual and weight transfer models.

Using the above model parameters, the major hyper-parameters, namely, the number of hidden layers and the number of neurons per hidden layer are tuned for both unilingual and multilingual TDNN-HMM models, as shown in Figure 4 and Figure 5, respectively. Figure 4 shows that the optimal number of TDNN layers which gives better performance for both unilingual and multilingual TDNN-HMM models is 8 with batch normalized ReLU hidden layers. The number of hidden layer seems large for training the unilingual TDNN-HMM models using the Chaha in-domain dataset but the ReLU nonlinearity and the dropout regularization enable us to increase the number of hidden layers without overfitting challenge. Therefore, according to the preliminary experimental results both the unilingual and multilingual TDNN-HMM models used the same number of hidden layers. The optimal number of neurons per hidden layer is experimented by making the number of hidden layers fixed to 8, and the results are presented in Figure 5. Figure 5 demonstrates that the optimal number of neurons per hidden layer is 450 for both unilingual and multilingual TDNN-HMM models.

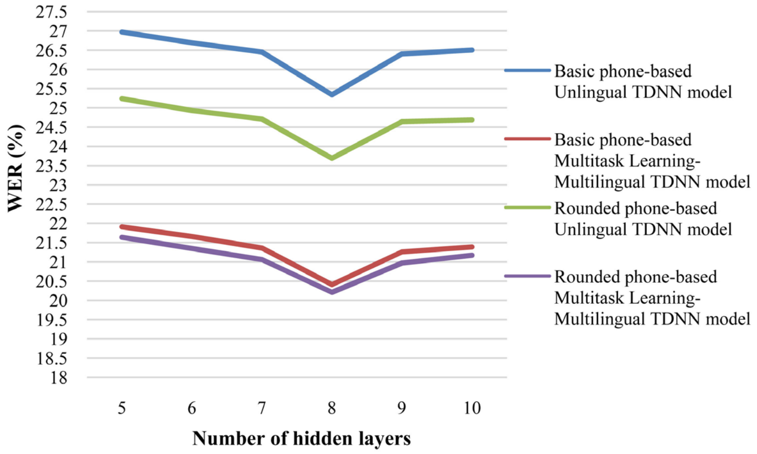

Figure 4. WER results vs. the number of hidden layers for basic phone and rounded phone-based unilingual and multilingual TDNN-HMM models.

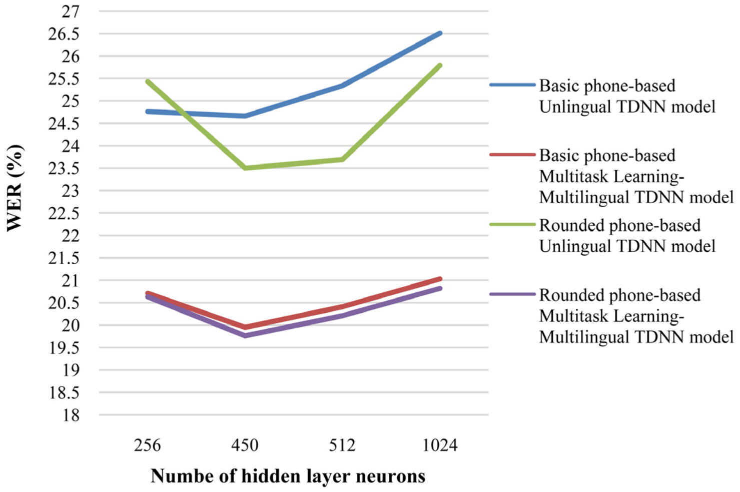

Figure 5. WER results vs. the number of neurons per hidden layer for basic phone and rounded phone-based unilingual and multilingual TDNN-HMM models.

5.2. Experimental Results and Discussion

5.2.1. Baseline GMM-HMM Models

Two baseline GMM-HMM models are trained, namely, GMM-CH, which trains using the Chaha language in-domain dataset, and GMM-MUL which is a phone sharing model that trains using the mixed multilingual data (Chaha and Amharic languages) by concatenating the Chaha and Amharic phone sets with a language identification. These models are developed for all basic and rounded triphone acoustic units.

Several GMM-CH models are trained for deciding the number of states and HMM topology, number of HMM leaves, and Gaussians for all basic and rounded phone acoustic modeling units. Hence, the standard 3-state with the fourth last non-emitting state Bakis HMM topology, and the 3-state with the fourth last non-emitting state with skip from the first state to the last non-emitting state Bakis HMM topology are empirically examined for all the basic and rounded phone acoustic modeling units, as shown in **Table 4**.

Table 4 indicates that the best HMM topology for both basic and rounded phone units-based GMM-CH model is a 3-state with the fourth last non-emitting state with skip from the first state to the last non-emitting state Bakis HMM topology. Using this topology, the best performing GMM-CH model that has a small word error rate (WER) consists of 1200 leaves *i.e.* 952 senones with 8000 Gaussians for the 36 phones of basic phone acoustic modeling units and 1200 leaves *i.e.* 960 senones with 8000 Gaussians for 41 phones of rounded phone acoustic modeling units. The results demonstrate that the rounded phone unit-based model outperforms the basic phone unit-based model with a relative WER reduction of 1.88%. All the rest GMM and DNN models are developed using 36 phones of basic phone units and 41phones of rounded phone units.

For training GMM-MUL models, the multilingual phone set is created by simply concatenating the Chaha and Amharic phone sets with a language identification prefix to ensure that the phones are distinct between languages. Fortunately, all phones of Amharic are similar with Chaha except the four palatalized phones, which are distinct for the Chaha language. Next, the GMM-MUL models are trained using 2432 and 2464 senones for basic and rounded phone acoustic units, respectively. Table 5 shows that both basic and rounded phone unit-based GMM-MUL models are worse in performance than the corresponding GMM-CH models.

Table 4. Number of HMM leaves and Gaussians, and the WER (%) of several GMM-CH models trained with 3-state HMM topologies.

Number of HMM leaves	Number of Gaussians	Transition Topologies	Acoustic modeling units (%WER)			
			Basic phone units		Rounded phone units	
			32 phones	36 phones	41 phones	44 phones
1000	6670	Standard	27.37	28.53	29.01	28.34
		With skip	28.72	27.68	27.47	26.89
1200	8000	Standard	29.11	27.27	27.56	27.85
		With skip	27.76	**26.02**	**25.53**	27.76
1400	9340	Standard	28.53	29.01	25.73	28.43
		With skip	28.05	27.76	26.50	26.60
1600	10670	Standard	28.92	27.56	28.34	27.47
		With skip	27.56	28.63	27.66	27.18

Table 5. WERs (%) of the baseline GMM-HMM models.

Model	Basic phone units	Rounded phone units
GMM-CH	26.02	25.53
GMM-MUL	27.95	27.66

5.2.2. DNN-HMM Models

Using the optimal parameters stated in Section 5.1, two unilingual TDNN models, namely, TDNN-CH and TDNN-AM are developed for Chaha and Amharic languages, respectively. The TDNN-AM models are used as the bootstrap to train the weight transfer multilingual models for the Chaha language. The TDNN-CH models are trained using the combined Chaha in domain real and synthetic speech corpus, and the results demonstrate that the rounded phone unit-based TDNN-CH model outperforms the equivalent basic phone unit-based model with an absolute WER reduction of 1.16%, as presented in the first row of Table 6. The finding shows that the use of rounded phones within lexical dictionary and phone list is improved the performance of Chaha ASR system.

Moreover, the effect of synthetic training dataset on the performance of basic and rounded phone unit-based TDNN models is examined by developing DNN-CH-Naug (None-augmented version of TDNN-CH) models using Chaha real training dataset (2.67 hrs), and compared with the TDNN-CH models developed using the Chaha total datasets (8.01 hrs). The basic phone and rounded phone unit-based TDNN-CH models achieved superior performance over the corresponding TDNN-CH-Naug models with absolute performance improvements of 6.19% and 4.84%, respectively, as presented in Table 6. Hence, augmenting the training dataset by generating the synthetic data using audio data augmentation technique improves the performances of ASR system for very low-resource languages.

We investigated three multilingual DNN models, explicitly, phone sharing (TDNN-MUL), multitask learning (TDNN-MT), and weight transfer models. The TDNN-MUL models are trained over the baseline GMM-MUL models, and realized superior performances over the unilingual TDNN-CH models with relative WER reductions of 18.82% and 13.57% for basic and rounded phone units, respectively, as presented in the second row of Table 7. Moreover, the basic phone unit-based TDNN-MUL model outperformed the corresponding rounded phone unit-based model with a relative WER reduction of 1.43%.

The TDNN-MT models are trained by sharing the hidden layers for Chaha and Amharic languages, and by making the output layer specific to each language. These models gain better performances than the corresponding unilingual TDNN-CH models with relative WER reductions of 19.10% and 15.91% for basic and rounded phone units, respectively, as presented in the third rows of Table 7. The rounded phone unit-based TDNN-MT model performed better

Table 6. WERs (%) of unilingual DNN-HMM models.

Model	Basic phone units	Rounded phone units
TDNN-CH	24.66	23.50
TDNN-CH-Naug	30.85	28.34
Absolute WER Reduction (%)	6.19	4.84

Table 7. WERs (%) of DNN-HMM models.

Model	Basic phone units	Rounded phone units
TDNN-CH	24.66	23.50
TDNN-MUL	20.02	20.31
TDNN-MT	19.95	19.76
TDNN-AM-WT	23.31	22.15
TDNN-MUL-WT	19.92	20.21
TDNN-MT-WT	**19.76**	**19.56**
Best Case Relative WER Reduction (%)	19.87	16.77

than the corresponding basic phone unit-based model with a relative WER reduction of 0.95%.

We investigated three weight transfer models, namely, weight transfer over TDNN-AM (TDNN-AM-WT), weight transfer over TDNN-MUL (TDNN-MUL-WT), and weight transfer over TDNN-MT (TDNN-MT-WT). To train these models, the seed TDNN-AM, TDNN-MUL, and TDNN-MT models are trained using Amharic, merged and shared multilingual (Amharic and Chaha datasets), respectively. Then, the last two hidden and output layers of the seed models are discarded and replaced by the new single hidden layer with 450 nodes, and the output layer with the number of output nodes equal to the number of senones, which are 656 and 696 for basic and rounded phone units, respectively. All the weights that connect the nodes of sixth hidden layer and biases are randomly initialized, and all the transferred hidden layers are fixed, and only the added hidden and output layers are trained using the Chaha training dataset.

The experimental results are presented from row 4 to 6 of Table 7, and all the weight transfer models improve the performances of the corresponding unilingual TDNN-CH models. Both the basic and rounded phone unit-based TDNN-MT-WT and TDNN-MUL-WT models realized superior performances over the corresponding TDNN-AM-WT model. This finding is because the performances of weight transfer models of Chaha are improved well, when the seed models are trained using both the Amharic and Chaha datasets. The rounded phone unit-based TDNN-MT-WT model is the best performing model with a relative WER reduction of 16.77% and it is also outperformed the corresponding TDNN-AM-WT model with a relative performance improvement of 1.01%.

5.2.3. Comparison of DNN-HMM Models Based on Their Performances

Table 7 reveals the following experimental results. All the basic and rounded phone unit-based multilingual TDNN models, namely, phone sharing, multitask learning, and weight transfer models outperform the baseline unilingual TDNN models consistently. This is because the unilingual TDNN models are trained using small training dataset than the multilingual TDNN models.

The multitask learning and phone sharing multilingual TDNN models realized better performances than the weight transfer multilingual TDNN models, when the seed models are trained using only the Amharic dataset. However, the performances of weight transfer multilingual TDNN models outperformed the multitask learning and phone sharing multilingual TDNN models if the seed models are trained using both Amharic and Chaha datasets. Among the multilingual TDNN models, TDNN-MT-WT model is the best performing model with best-case relative WER reductions of 19.87% and 16.77% for basic and rounded phone acoustic modeling units, respectively.

The rounded phone unit-based TDNN-CH, TDNN-MT, TDNN-AM-WT, and TDNN-MT-WT models outperformed the corresponding basic phone unit-based models with relative WER reductions of 4.7%, 0.95%, 4.98%, and 1.01%, respectively. This is because the characteristics of the language, when considering the rounded nature of the rounded phones during acoustic modeling, the performances of the unilingual and multilingual TDNN models are improved. Hence, the rounded phone units are the best acoustic modeling units to develop reliable Chaha ASR system.

On the other hand, the rounded phone unit-based TDNN-MUL and TDNN-MUL-WT models are worse in performance than the equivalent basic phone unit-based models with relative WER reductions of 1.43%. Overall, most of the rounded phone unit-based unilingual and multilingual TDNN models outperformed the equivalent basic phone unit-based models. However, the performances of both acoustic unit-based models are comparable to each other. Hence, the basic phone units can be used as alternative acoustic modeling units to develop Chaha ASR system.

5.2.4. Comparison of DNN-HMM Models Based on Their Recognition Speeds

The speed of a speech recognition system is measured using a real time factor (RTF). RTF is a very natural measure of a speech decoding speed which expresses how much the speech recognition system decodes slower than the user speaks. It is the ratio of the speech recognition system response time to the utterance duration, as formulated in Equation (1).

$$\text{RFT} = \frac{\text{time}\left(\text{decode}\left(\text{a}\right)\right)}{\text{length}\left(\text{a}\right)} \tag{1}$$

where a is an utterance. Usually both the average RTF (average over all utterances) and 90th percentile RTF is examined in efficiency analysis of speech recognition system. We have used an average RTF of all the utterances to analysis the speed of all the speech recognition systems developed in this study. Hence, the recognition speeds of the basic and rounded phone unit-based unilingual and multilingual TDNN models are presented in Table 8. Both the basic and rounded phone unit-based unilingual TDNN models are faster than the equivalent phone sharing and multitask learning multilingual TDNN models. Likewise,

Table 8. Recognition speeds of DNN-HMM models.

Model	Real Time Factor	
	Basic phone units	Rounded phone units
TDNN-CH	0.225	0.206
TDNN-MUL	0.278	0.283
TDNN-MT	0.284	0.279
TDNN-AM-WT	**0.086**	**0.086**
TDNN-MUL-WT	0.088	**0.086**
TDNN-MT-WT	0.088	**0.086**

the basic and rounded phone unit-based unilingual, phone sharing, and multitask learning TDNN models are slower than the corresponding weight transfer multilingual TDNN models. Hence, the fastest recognition speeds are realized using the basic phone unit-based TDNN-AM-WT, and rounded phone unit-based TDNN-AM-WT, TDNN-MUL-WT and TDNN-MT-WT models with real-time factor of 0.086. Almost all the rounded phone unit-based multilingual TDNN models are faster than the corresponding basic phone unit-based models. The reason for this is because making all the decoding parameters universal for all models, the recognition speed varies with graph size, and thus the graph size of all the rounded phone unit-based multilingual TDNN models are smaller than the corresponding basic phone unit-based models. The graph sizes of the basic and rounded phone unit-based multilingual TDNN models increase because the size of the training dataset is increased to realize better performance, and this leads to relatively slow recognition speeds.

Overall, the performances of rounded phone unit-based multilingual TDNN models are better than the corresponding basic phone unit-based models, as discussed in Section 5.2.3. In line with this, the recognition speeds of basic phone unit-based multilingual TDNN models are worse than the corresponding rounded phone unit-based models. Hence, the rounded phone units are the best acoustic modeling units to develop ASR system for the very low-resource language, Chaha.

6. Conclusions and Future Works

This study presents a first attempt made on the investigation of ASR systems using various multilingual DNN techniques for the very low-resource language, Chaha. The language and resource-related problems are the major factors that challenge the development of Chaha ASR systems. By considering these challenges, this paper investigated different unilingual and multilingual speech recognizers. The experimental results demonstrate that all the basic and rounded phone unit-based multilingual TDNN models realized superior performances over the corresponding unilingual TDNN models with overall relative WER reductions of 5.47% to 10.87% and 5.74% to 16.77%, respectively. Hence, multilingual DNN models are profoundly effective and efficient to develop better performance speech recognizers for the very low-resource languages, which are spoken in the developing and minority countries. Moreover, both basic and rounded phone unit-based multilingual TDNN models achieved comparable recognition performances and decoding speeds. Hence, both basic and rounded phone acoustic modeling units are convenient to develop ASR system for Chaha. However, almost all the rounded phone unit-based unilingual and multilingual models realized superior performances and faster recognition speeds than the corresponding basic phone unit-based models. Hence, the rounded phone units are the best acoustic modeling units to develop reliable ASR system for the very low-resource language, Chaha.

As future work, we are interested in exploring the use of CV syllables as acoustic modeling units for building Chaha ASR system. Besides, the language-specific issues like gemination and devoicing of consonants, proper insertion of an epenthetic vowel, and pronunciation of open vowels in the training corpus and pronunciation dictionaries will need to be handled.

Acknowledgements

We gratefully acknowledge the granted financial support from the National Natural Science Foundation of China (No. 61572211, 61173114, 61202300). We also thank Mr. Girum Dagnaw for his proofreading our manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Besacier, L., Barnard, E., Karpov, A. and Schultz, T. (2014) Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication*, **56**, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

[2] Vu, N.T., Imseng, D., Povey, D., Motlícek, P., Schultz, T. and Bourlard, H. (2014) Multilingual Deep Neural Network Based Acoustic Modeling for Rapid Language Adaptation. *Proceedings of International Conference on Acoustics*, *Speech and Signal Processing*, Florence, 4-9 May 2014, 7639-7643. https://doi.org/10.1109/ICASSP.2014.6855086

[3] Chen, D. and Mak, B.K.-W. (2015) Multi-Task Learning of Deep Neural Networks for Low-Resource Speech Recognition. *IEEE*/*ACM Transactions on Audio*, *Speech*, *and Language Processing*, **23**, 1172-1183. https://doi.org/10.1109/TASLP.2015.2422573

[4] Ghahremani, P., Manohar, V., Hadian, H., Povey, D. and Khudanpur, S. (2017) Investigation of Transfer Learning for ASR Using LF-MMI Trained Neural Networks. *Proceedings of Automatic Speech Recognition and Understanding*, Okinawa, 16-20 December 2017, 279-286. https://doi.org/10.1109/ASRU.2017.8268947

[5] Huang, J., Li, J., Yu, D., Deng, L. and Gong, Y. (2013) Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network with Shared Hidden Layers.

Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vancouver, 26-31 May 2013, 7304-7308. https://doi.org/10.1109/ICASSP.2013.6639081

[6] Gales, M.J., Knill, K., Ragni, A. and Rath, S.P. (2014) Speech Recognition and Keyword Spotting for Low-Resource Languages: Babel Project Research at CUED. In: *SLTU*, ISCA, St Petersburg, 16-23.

[7] Lin, C., Wang, Y., Chen, S. and Liao, Y. (2016) A Preliminary Study on Cross-Language Knowledge Transfer for Low-Resource Taiwanese Mandarin ASR. *Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques*, Bali, 26-28 October 2016, 33-38. https://doi.org/10.1109/ICSDA.2016.7918980

[8] Miao, Y. and Metze, F. (2013) Improving Low-Resource CD-DNN-HMM Using Dropout and Multilingual DNN Training. *Proceedings of Interspeech*, Lyon, 2237-2241.

[9] Imed, Z. (2014) Natural Language Processing of Semitic Languages. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45358-8

[10] Abate, S.T. and Menzel, W. (2007) Automatic Speech Recognition for an Under-Resourced Language—Amharic. *Proceedings of Interspeech*, Antwerp, 1541-1544.

[11] Tachbelie, M.Y., Abate, S.T., Besacier, L. and Rossato, S. (2012) Syllable-Based and Hybrid Acoustic Models for Amharic Speech Recognition. *Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape Town, 7-9 May 2012, 5-10.

[12] Tachbelie, M.T., Abate, S.T. 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. https://doi.org/10.1016/j.specom.2013.01.008

[13] Leslau, W. (1997) Chaha (Gurage) Phonology. In: Kaye, A.S., Ed., *Phonologies of Asia and Africa*, Vol. 2, Eisenbrauns, Winona Lake, 373-397.

[14] Banksira, D.P. (2000) Sound Mutations: The Morphophonology of Chaha. John Benjamins Publ., Amsterdam. https://doi.org/10.1075/z.93

[15] Carolyn, M.F. (1986) Notes on the Phonology and Grammar of Chaha-Gurage. *Journal of Ethiopian Studies*, **19**, 41-80.

[16] Rose, S. (2000) Epenthesis Positioning and Syllable Contact in Chaha. *Phonology*, **17**, 397-425. https://doi.org/10.1017/S0952675701003931

[17] Rose, S. (2007) Chaha (Gurage) Morphology. In: Kaye, A.S., Ed., *Morphologies of Asia and Africa*, Eisenbrauns, Winona Lake, 399-424.

[18] Lewis, M.P. (2009) Ethnologue: Languages of the World. Sixteenth Edition, SIL International, Dallas. http://www.ethnologue.com/16

[19] Besacier, L., Le, V.-B., Boitet, C. and Berment, V. (2006) ASR and Translation for Under-Resourced Languages. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, 14-19 May 2006, 1221-1224. https://doi.org/10.1109/ICASSP.2006.1661502

[20] Abate, S.T., Menzel, W. and Tafila, B. (2005) An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition. *Proceedings of Interspeech*, Lisbon, 1601-1604.

[21] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S. (2015) Audio Augmentation for Speech Recognition. *Proceedings of Interspeech*, Dresden, 6-10 September 2015,

3586-3589.

[22] 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. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, 11-15 December 2011.

[23] Stolcke, A. (2002) SRILM—An Extensible Language-Modeling Toolkit. *Proceedings of International Conference on Spoken Language Processing*, Denver, 901-904.

[24] Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y. and Khudanpur, S. (2016) Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. *Proceedings of Interspeech*, San Francisco, 8-12 September 2016, 2751-2755. https://doi.org/10.21437/Interspeech.2016-595