Received June 28, 2019, accepted July 10, 2019, date of publication July 26, 2019, date of current version August 15, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2931391*

# Investigation of Various Hybrid Acoustic Modeling Units via a Multitask Learning and Deep Neural Network Technique for LVCSR of the Low-Resource Language, Amharic

**TESSFU GETEYE FANTAYE[1], JUNQING YU[1,2], AND TULU TILAHUN HAILU[1]**
[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Center of Network and Computation, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Junqing Yu (yjqing@mail.hust.edu.cn)

This work was supported by National Natural Science foundation of China under grant numbers of 61572211, 61173114, and 61202300.

**ABSTRACT** Multitask learning (MTL) is helpful for improving the performance of related tasks when the training dataset is limited and sparse, especially for low-resource languages. Amharic is a low-resource language and suffers from the problems of training data scarcity, sparsity, and unevenness. Consequently, fundamental acoustic units-based speech recognizers perform worse compared with the speech recognizers of technologically favored languages. This paper presents the results of our contributions to the use of various hybrid acoustic modeling units for the Amharic language. The fundamental acoustic units, namely, syllable, phone, and rounded phone units-based deep neural network (DNN) models have been developed. Various hybrid acoustic units have been investigated by jointly training the fundamental acoustic units via the MTL technique. Those hybrid units and the fundamental units are discussed and compared. The experimental results demonstrate that all the fundamental units-based DNN models outperform the Gaussian mixture models (GMM) with relative performance improvements of 14.14%–23.31%. All the hybrid units outperform the fundamental acoustic units with relative performance improvements of 1.33%–4.27%. The syllable and phone units exhibit higher performance under sufficient and limited training datasets, respectively. All the hybrid units are useful with both sufficient and limited training datasets and outperformed the fundamental units. Overall, our results show that DNN is an effective acoustic modeling technique for the Amharic language. The context-dependent (CD) syllable is the more suitable unit if a sufficient training corpus is available and the accuracy of the recognizer is prioritized. The CD phone is a superior unit if the available training dataset is limited and realizes the highest accuracy and fast recognition speed. The hybrid acoustic units perform the best under both sufficient and limited training datasets and achieve the highest accuracy.

**INDEX TERMS** Acoustic modeling units, Amharic, hybrid acoustic modeling units, low-resource language, multitask learning.

## I. INTRODUCTION

Deep neural networks (DNNs) were introduced into speech recognition research in 2011 as an acoustic modeling technique in the hybrid DNN-Hidden Markov Model (HMM) and as a feature extractor for the tandem GMM-HMM and DNN-HMM models. Subsequently, many studies were conducted to explore the effectiveness of DNNs for context-independent and context-dependent large-vocabulary continuous speech recognition (LVCSR) tasks. These studies were conducted on both high-resource and low-resource languages using large, very large, and small training datasets. The studies that were conducted on high-resource languages include [1]–[7]. All these works were conducted using the

The associate editor coordinating the review of this manuscript and approving it for publication was Stavros Ntalampiras.

hybrid DNN-HMM model, in which a DNN is used to compute the posterior probabilities of the context-independent states and context-dependent tied states while the HMM is used to model the temporal-sequential characteristics of the speech.

Previous studies such as [8]–[17] investigated low-resource languages using a DNN as the hybrid in the monolingual, multilingual, multitasking, and cross-language model transfer approaches and as the feature extractor for the tandem and DNN-HMM models. All the studies on both high-resource and low-resource languages concluded that the DNN outperforms the traditional GMM acoustic model in terms of a significant word error rate (WER) reduction. As a result, at this time, the deep neural network and its various architectures have become the state-of-the-art acoustic modeling techniques for LVCSR tasks. However, the hybrid DNN-HMM models of low-resource languages are outperformed by the corresponding high-resource language models. This phenomenon is because low-resource languages suffer from the problems of training data scarcity, sparsity and unevenness. As a result, the hybrid DNN-HMM models tend to overfit. The overfitting problem is mitigated via various approaches, such as applying model size optimization parameters such as weight regularization (e.g., dropout) and activation functions (e.g., maxout, pnorm, and softmaxout) [18]–[20] and using sufficient training datasets from high-resource languages via various training data sharing paradigms such as multilingual learning [12], [21], [22], cross-language model transfer [14], and multitask learning [10], [11].

Multitask learning is successfully used to training related languages using training datasets from diverse languages as shared hidden-layer multilingual models for sharing knowledge among the languages [14], [23], [24]. Similarly, multitask learning is also used to jointly train related tasks (with different acoustic modeling units) using only a single-language dataset, without the need for additional datasets from other languages [10], [11], [25]–[28]. In MTL, if the tasks are related, learning them together can facilitate the transfer of knowledge among tasks because it effectively increases the amount of training data for each task. Hence, MTL is typically helpful when the training data size is small compared to the model size. Thus, this learning paradigm is useful for low-resource-language speech recognition tasks.

Triphones and senones are the standard acoustic modeling units that are used in the hybrid DNN-HMM systems for automatic speech recognition. However, these acoustic units cannot exploit both spectral and temporal characteristic of continuous speech. Hence, longer acoustic modeling units, such as syllables and graphemes, are used to overcome the limitations of the context-dependent phones and senones units. Nevertheless, longer-unit-based acoustic models are profoundly affected by the problems of data scarcity, sparsity, and unevenness, especially for low-resource languages. To overcome these challenges, the hybrid acoustic modeling units are used either to manipulate the training dataset with backoff sparse long acoustic units to the context-dependent phone units [29] or to jointly train fundamental acoustic modeling units as a hybrid modeling unit using the MTL-DNN approach [4], [10], [11], [26].

Amharic has two fundamental acoustic modeling units, namely, phones (basic and rounded phones) and syllables (vowel (V), consonant-vowel (CV), vowel-consonant (VC), consonant-vowel-consonant (CVC), vowel-consonant-consonant (VCC), and consonant-vowel-consonant-consonant (CVCC)), which are used to develop LVCSR systems. Amharic is also a low-resource language and has an insufficient training corpus. This insufficiency affects the performance of speech recognizers that are developed using the above basic acoustic modeling units. As a result, the phone-based acoustic models suffer from the problem of data scarcity and syllable-based acoustic models suffer from the problems of data scarcity, sparsity and unevenness. These problems are mitigated by either preparing a sufficient training corpus or building an acoustic model using the existing training datasets via the various behavior DNN models. However, the development of a training corpus requires and consumes much more time, human labor, and financial resources. Thus, this study examines hybrid acoustic modeling units via a MTL-DNN scheme for building an acoustic model for the Amharic language using the available training data. This approach is selected due to the performance of the DNN and MTL acoustic modeling schemes and the relatedness of the basic acoustic modeling units of the language. First, the DNN can extrapolate new features from a limited set of features that are contained in a training set, and it has learnable activation functions via which the data are mapped. The underlying data distribution is approximated for the sparse dataset. This approach uses the same hidden layers for related tasks when it is used for joint training. Hence, the DNN reduces the data sparsity and unevenness in the limited training datasets. Second, MTL is especially useful if the size of the training dataset is limited and if there are sparse and unevenly distributed acoustic modeling units in the limited training data. Third, the language fundamental acoustic modeling units (CD syllable, CI syllables, CD phones, CI phones, CD rounded phones, and CI rounded phones) are related tasks, and they can be trained using the same acoustic input features.

This paper offers the following contributions:

- Exploring the use of the DNN acoustic modeling technique to build a LVCSR system for Amharic by developing the optimal DNN acoustic models using the fundamental acoustic modeling units, namely, the CD syllable, CI syllable, CD phone,[1] CI phone, CD rounded[2] phone, and CI rounded phone.

[1] Phone acoustic units contain only basic phones, where the rounded phones map to the corresponding basic phones.

[2] Rounded phone acoustic 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.

- Analyzing the influence of data sparsity on the performance of syllable-based models and comparing the speaker-independent and speaker-adapted DNN models to minimize the data sparsity and uneven distribution of syllables in the existing limited datasets.
- Proposing various new hybrid[3] acoustic modeling units by training the fundamental acoustic modeling units jointly via the MTL-DNN scheme for Amharic language.
- Comparing the proposed hybrid acoustic modeling units with the fundamental acoustic units in terms of the recognition performance and speed and suggesting when to apply these units to develop a higher performing speech recognizer for Amharic language.
- Developing and evaluating the proposed hybrid unit and fundamental acoustic unit-based speech recognizers using the dataset from [30] and own, and the hybrid unit-based recognizers are obtained superior performances and recognition speeds over the fundamental acoustic unit-based recognizers.

The remainder of this paper is organized as follows. Section II presents the MTL-DNN acoustic modeling paradigm. A description of the Amharic language and the acoustic modeling units of the Amharic speech recognition system are presented in Section III. Section IV explains the corpora that are used for training and testing the models. The experiments, results, and discussion are presented in Section V. The conclusions and future directions of this work are discussed in Section VI.

## II. MTL-DNN ACOUSTIC MODELING PARADIGM

Multitask learning is a machine learning strategy that is used to improve the overall performance of a learning task by jointly learning multiple associated tasks [11]. Multitask learning helps to transfer knowledge between or among tasks if the tasks are associated with each other and share an internal representation by joint learning [11]. In this approach, most of the DNN learns a primary task, in addition to one, two or more ancillary tasks, and these ancillary tasks aim to help the model converge to the benefit of the primary task. The selection of ancillary tasks has a significant impact on the performance of the primary task. If the ancillary tasks are well selected, the primary task can improve its robustness to unseen data, thereby leading to better generalization. The key benefits of MTL for the DNN include model regularization, attention focusing, eavesdropping, representation bias, and implicit data augmentation [31].

MTL has been effectively used as the shared hidden layer in multilingual modeling to train related languages for sharing knowledge among them [14], [23], [24]. MTL is an efficient method for low-resource-language speech recognition because it profoundly improves the performance on the individual languages that have been jointly trained. MTL has also been used to jointly train related acoustic modeling units using the specified language dataset without the need for additional training datasets from other languages [4], [10], [11], [25]–[28]. Seltzer and Droppo have investigated the joint training of monophone state posteriors as a primary task with state context, phone context or phone labeling as ancillary tasks to improve the phoneme recognition performance [28]. Chen *et al.* [10] were motivated by [28] to examine the MTL-DNN model for the joint training of triphone and tri-grapheme acoustic models for under-resourced-language speech recognition, where the triphones were used as the main task and the tri-graphemes were used as an ancillary task [10]. Chen and Mak [27] have explored the joint training of the monophone and senones as ancillary tasks with the distinct triphone states for TIMIT and WSJ phone speech recognition. Chen and Mak [11] have also analyzed the training of language-specific triphones and a universal phone set for various languages via the MTL-DNN for low-resource languages. Li *et al.* have explored hybrid acoustic units that combine the context-dependent phones as the main task with context-dependent initials/finals and syllable acoustic units as auxiliary tasks using MTL for Chinese speech recognition [4]. Bell and Renals have proposed a monophone classification secondary task for CD phone acoustic modeling [25]. Bell *et al.* have also demonstrated the joint training of context-dependent phones as a primary task with context-independent phones as an ancillary task, and the context-dependent phones train with senones as an auxiliary task with context-dependent phones as a primary task in the MTL approach [26]. All these studies achieve superior performance compared to the equivalent single-task DNN models. These studies demonstrate the benefits of MTL for speech recognition tasks in general and for low-resource-language speech recognition tasks in particular.

Amharic is a low-resource language with the limitations of data sparsity and unevenly distributed acoustic units in the available training corpus. These limitations lead to inferior speech recognizer performance compared with high-resource languages. To overcome these limitations and to improve the recognizer performance, we have explored new hybrid acoustic modeling units by jointly learning the fundamental acoustic units of the language in the two-task and three-task learning paradigms.

## III. AMHARIC LANGUAGE AND ACOUSTIC MODELING UNITS OF THE AMHARIC LVCSR SYSTEM

### A. THE AMHARIC LANGUAGE

The Amharic language is the working language of the Federal Democratic Republic of Ethiopia. This language, which is named Amarinya or Amarigna, is the second most widely spoken language in the semantic language family, after Arabic. Based on the 2007 census, Amharic has over 22 million native speakers in Ethiopia [32], [33]. There are 2.7 million additional speakers who live in other countries

[3] Hybrid acoustic modeling unit refers to the joint training of either two or three basic acoustic modeling units via MTL.

**TABLE 1. Amharic consonants (adapted from [34]).**

| Manner of articulation | | Place of articulation | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Labials | | Alveolar | | Palatals | | Velars | | Labiovelar | | Glottals | |
| Stops | Voiceless | p | ፕ | t | ት | | | k | ክ | $k^w$ | ኰ | ? | ዕ |
| | Voiced | b | ብ | d | ድ | | | g | ግ | $g^w$ | ጐ | | |
| | Glottalized | p’ | ጵ | t’ | ጥ | | | q | ቅ | $q^w$ | ቈ | | |
| | Rounded | | | | | | | | | | | h | ህ |
| Fricatives | Voiceless | f | ፍ | s | ስ | š | ሽ | | | | | | |
| | Voiced | v | ቭ | z | ዝ | ž | ዥ | | | | | | |
| | Glottalized | | | s’ | ጽ | | | | | | | $h^w$ | ኈ |
| | Rounded | | | | | | | | | | | | |
| Affricative | Voiceless | | | | | č | ች | | | | | | |
| | Voiced | | | | | ğ | ጅ | | | | | | |
| | Glottalized | | | | | č’ | ጭ | | | | | | |
| | Rounded | | | | | | | | | | | | |
| Nasals | Voiced | m | ም | n | ን | ň | ኝ | | | | | | |
| Liquid | Voiced | | | l | ል | | | | | | | | |
| | | | | r | ር | | | | | | | | |
| Glides | | w | ው | | | y | ይ | | | | | | |

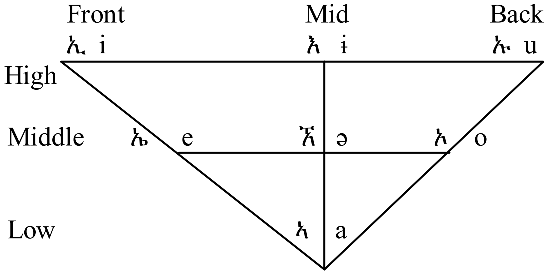

**FIGURE 1. Amharic vowels with their features (adapted from [34]).**

such as Israel, Eretria, Canada, United States of America, Germany, and Sweden [35]. The language has five dialects: Addis Ababa, Gojam, Gonder, Wollo, and Menz [36].

The phonetic set of the Amharic language consists of 7 vowels and 32 consonants, which yield 39 phonemes; these phonemes constitute the complete inventory of sound units of the language [34]. The seven vowels are ə, u, i, a, e, ɨ, and o. These vowels can be classified as rounded and unrounded; as front, middle, and back; and as high, middle, and low based on the lip rounding and horizontal and vertical tongue movements, respectively, as shown in Figure 1. The 32 consonants are classified as stops, fricatives, affricatives, nasals, liquids, and semivowels based on the manner of articulation, as listed in Table 1.

The Amharic writing script is syllabic, where each character represents a combination of a consonant with a vowel to compose a CV syllabic structure, except for the glottalized and sixth-order consonants [29], [37]. The glottalized and sixth-order consonants can be pronounced with or without vowels. As a result, the language has a total of 276 distinctive writing symbols, where 231 graphemes of 33 core symbols with seven orders as listed in Table 2, 20 graphemes of four labiovelar symbols with five orders, 18 graphemes of eighteen labialized symbols with one order, and 7 graphemes of one labiodental symbol with seven orders [29], [37].

**TABLE 2. Sample core letters of Amharic.**

| Consonants (initials) | Vowels(finals) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1st<br>ə | 2nd<br>u | 3rd<br>i | 4th<br>a | 5th<br>e | 6th<br>ɨ | 7th<br>o |
| h | ሀ | ሁ | ሂ | ሃ | ሄ | ህ | ሆ |
| l | ለ | ሉ | ሊ | ላ | ሌ | ል | ሎ |
| | | | … | | | | |
| f | ፈ | ፉ | ፊ | ፋ | ፌ | ፍ | ፎ |
| p | ፐ | ፑ | ፒ | ፓ | ፔ | ፕ | ፖ |

Moreover, Amharic is a syllabic language that has V, VC, VCC, CV, CVC, and CVCC possible syllable structures [34], [37]. Of these syllables, the CV syllable is the most widely distributed syllable in the language; 231 core and 7 labiodental letters are basic CV syllables, while the 20 labiovelar and 18 labialized letters are combinations of two or three CV syllables. There are syllables that have different symbolic representations and similar pronunciations. These syllables are (ሀ/h/, ኸ/h/, ሐ/h/, and ኀ/h/), (ሰ/s/ and ሠ/s/), (ፀ/t’/ and ጸ/t’/), and (ዐ/? / and አ/? /).

However, the speech recognition task requires a unique pronunciation of each orthographic syllable symbol. Hence, such syllables should be reduced to the common syllables.

The common syllables that are used as the reduced syllables are ሀ/h/, ሰ/s/, ጸ/t'/, and አ /ʔ/ for (ሀ, ኸ, ሐ, and ኀ), (ሰ and ሠ), (ፀ and ጸ), and (ዐ and አ), respectively. As a result, six syllables are removed, and each of these syllables has seven orders; hence, a total of 42 syllables are removed (6*7 orders = 42 syllables). The 276 syllables are reduced to 234 syllables. However, of these 234 syllables, there are two syllables that have similar pronunciations in the first and fourth syllable orders: (አ/a/ and ኣ/a/) and (ሀ/h/ and ሃ/h/). Of these syllables, we have used the first-order syllable for our study. Thus, 232 syllables remain. However, syllable ኸ/hə/ has a distinct pronunciation, and we have considered it as a uniquely pronounced syllable. Hence, we have used a total of 233 unique CV syllables in our study [30].

### B. ACOUSTIC MODELING UNITS OF THE AMHARIC LVCSR SYSTEM

In the development of a speech recognition system, the utilization of different acoustic modeling units depends on the nature of the target language and the size of the available training dataset. Based on the characteristics of the language, Amharic has two major fundamental acoustic modeling units: phones and syllables. By using these acoustic modeling units, researchers have been attempting to develop speech recognition systems for Amharic since 2002. Tadesse [38] investigated the use of sub-word acoustic units, namely, phones (CI and CD) and CV syllables (CI and CD), for small-vocabulary isolated word recognition. Tadesse suggested that the use of CV syllables leads to low performance relative to phones in the acoustic modeling. Abate and Menzel, in [36] and [39], obtained similar results to those of Tadesse [38] when they developed speech recognition systems using the triphone and CI CV syllable as acoustic units. These authors reported that the CI CV syllable is outperformed by the triphone in terms of accuracy; however, in terms of recognition speed and required storage space, the CI CV syllable outperforms the triphone.

However, Tachbelie *et al.* [29] and Tachbelie *et al.* [37] found that the poor performance in Tadesse [38] was due to the use of limited training datasets. Those authors investigated the tied-state triphone, CD syllable, CI syllable, and hybrid (syllable and phone) acoustic units for Amharic LVCSR. According to these findings, the hybrid (phone-syllable)-based recognizers did not achieve significant performance improvements over the highest performing CD-syllable-based recognizers when word units were used in the pronunciation dictionary and language model. However, the recognizers yielded larger WER reductions in morpheme-based speech recognition. CD CV syllable units are the best alternatives if and only if a sufficient training dataset is available; if the training dataset is limited, the CI syllable units are more suitable than tied-state triphone acoustic units. Woldeyohannis *et al.* [40] recently investigated phoneme (triphone) and CV syllable acoustic units and demonstrated that the phoneme (triphone) acoustic units outperformed the CV syllable units.

Alternatively, Gebremedhin *et al.* [35] explored the common acoustic model for similarly pronounced CV syllables, which is based on the vowel articulation. Hence, the number of CV syllables that are used as acoustic modeling units was 93, which was reduced from 233 CV syllables. Dribssa and Tachbelie [41] investigated the use of various syllable types, namely, V, CV, VC, CVC, VCC, and CVCC, as acoustic units. These authors demonstrated that the use of all syllable types as acoustic units is promising for LVCSR if the training dataset is sufficiently large. In all prior works, the conventional GMM acoustic modeling approach was employed. Seid and Gambäck [42] examined the effectiveness of the artificial neural network (ANN)-HMM model using phone acoustic units and they concluded that the use of the ANN as an acoustic modeling technique yields promising results for the Amharic language.

In prior works, the researchers prepared their own training datasets for accomplishing their objectives since Amharic is a low-resource language. Abate *et al.* [30] developed a medium-sized corpus of 20 hours for the language. This corpus is small compared with other languages. Moreover, Amharic has 59,319 triphone and 12,649,337 tri-syllable modeling units for phone and syllable units, respectively. Hence, the modeling units of the phones are 213 times less numerous than those of the syllables. Hence, the phones are easy to model using a relatively small training dataset because the modeling units are small and the phones are distributed evenly throughout the training dataset. However, syllables are longer than phones since each syllable contains two or more phone units, and they have many modeling units. Hence, they require a large training dataset and evenly distributed syllables in the training dataset for building a reliable LVCSR system. As a result, syllable acoustic modeling units are challenged by the problems of data scarcity and sparsity. To overcome these challenges, this study proposes hybrid acoustic modeling units by jointly training the fundamental acoustic modeling units to share the training data among them via the MTL-DNN paradigm, in contrast with Tachbelie *et al.* [29] and Tachbelie *et al.* [37], in which the hybrid units are suggested by manipulating the training dataset via backoff from sparse syllables to phones.

## IV. PREPARATION OF THE CORPORA

### A. SPEECH CORPUS

As presented in Table 3, the training speech corpus that is used in our investigation is obtained from two sources. The first source is Abate *et al.* [30], which has 20-hour read training speech corpus that was collected from 50 male and 50 female speakers. The second source is a 6-hour training speech corpus, which we prepared from two radio news broadcasting corporations, namely, the Deutsche Welle and Voice of America Radio stations, and a total of 2,674 sentences (14,209 types or 44,133 tokens) are taken from 15 male and 10 female news readers. We have also used the 5k development test set from [30] as a testing dataset, which was collected from 20 speakers.

**TABLE 3.** Speech training corpus.

| Data Sources | Speakers | Sentences | Word Tokens | Duration (Hrs) |
|---|---|---|---|---|
| Solomon [30]-train | 100 | 10,875 | 28,666 | 20 |
| Solomon [30]-test | 20 | 359 | 4,106 | 1 |
| Own-train | 25 | 2,674 | 44,133 | 6 |
| Total | 145 | 13,908 | 76,905 | 27 |

However, the syllable distribution in the total training speech corpus is uneven, as presented in Table 4. Of the 233 uniquely pronounced CV syllables, 9.9%, 20.6%, 36.1%, and 42.9% of the syllables have frequencies of less than 50, 100, 300, and 500 in the training speech corpus. Moreover, compared to other speech corpora that contain hundreds or thousands of hours of speech data for training, this corpus is small; hence, the models will suffer due to lack of training data.

**TABLE 4.** Syllable frequency distribution in the training speech corpus.

| Frequency | Number of syllables |
|---|---|
| <100 | 48 |
| 100-499 | 52 |
| 500-999 | 33 |
| 1000-4999 | 64 |
| 5000-10000 | 18 |
| >10000 | 18 |

### B. TEXT CORPUS

The text corpus that is used in this study is obtained from two sources: The first source is a text corpus from Tachbelie and Abate [43], which contains 217,566 sentences, and the second source is an Amharic-English bilingual corpus from European Languages Resource Association (ELRA), which contains 6,613 news sentences [44]. However, the 217,566 sentences consist of many duplicate and testing dataset sentences, which are removed to avoid overlapping between the text corpus and the testing dataset, and 209,463 sentences are obtained. Then, these sentences are merged with the 6,613 news sentences and a total of 216,076 sentences (313,488 word types and 4,553,669 word tokens) are obtained; the resulting text corpus that mainly used to generate lexical dictionaries and to train language models.

In this study, the phone-level Unicode versions of the text corpus and transcribed speech text are used. Thus, transliteration[4] of the text corpus and the transcribed speech text from their syllable-level Unicode versions into the corresponding phone-level Unicode versions can be conducted as follows: All the syllables except the 20 labiovelars and 18 labialized syllables are transliterated in a CV manner. For example, the word በሬ/bera/, which means ‘ox’, is transliterated as ብኧርኤ/bəre/, where syllable በ/bə/ is transliterated as the combination of the sixth-order phone, namely, ብ/b/, with the first vowel, namely, ኧ/ə, to the transliterated form of ብኧ/bə/ and syllable ሬ/re/ is transliterated as the combination of sixth-order phone ር/r/ with the 5th vowel, namely, ኤ/e/, to the transliterated form of ርኤ/re/.

However, the labiovelar and labialized syllables are combinations of two or three CV syllables. Thus, according to [45]–[47], these syllables can be transliterated as the concatenations of sixth-order phones with rounded vowels. For instance, syllable ቋ/qwa/ is a labialized syllable and is transliterated as a combination of sixth-order phone ቅ/q/ with rounded vowel ውኣ/wa/ to the corresponding phone transliteration of ቅውኣ/qwa/. Similarly, አንኳኲ/ankwakwi/, which means “she knocks”, is transliterated as አ/a/ + ንእ/nɨ/ + ክውኣ/kwa/ + ክውኢ/kwi/, where the first term is a vowel, which is taken directly; the second term is a basic syllable, which is transliterated via the CV approach; and the third and the fourth terms are labialized and labiovelar syllables, which are transliterated as combinations of the sixth-order phones with rounded vowels.

### C. LEXICAL DICTIONARIES

According to the acoustic modeling units that are investigated, we construct three lexical dictionaries: syllable, phone, and rounded phone dictionaries. All three dictionaries are derived from the text corpus by selecting the most frequent words and are of size 85k. The syllable-based lexicon contains 233 uniquely pronounced CV syllables. The phone-based lexical dictionary contains 34 phones: seven vowels and 27 consonants. The rounded-phone-based lexicon considers the rounded nature of the rounded consonants (labiovelars and labialized) in addition to the basic 34 phones. Therefore, the lexicon contains a total of 39 phones: seven vowels, 27 consonants, and five rounded vowels. In all lexical dictionaries, language properties such as gemination, insertion of an epenthetic vowel, glottal stop consonant pronunciation, and elision of vowels or insertion of semivowels are not considered; this is because the lexicons are prepared by nonlinguistic experts with the help of the writing system of the language.

## V. EXPERIMENTATION

### A. EXPERIMENTAL SETUP

The baseline GMM-HMM models are developed using a state-of-the-art speech recognition toolkit, namely, Kaldi [48]. In the feature extraction process, 16-kHz speech input is coded with 13-dimensional Mel-frequency cepstral coefficients (MFCCs) with a 25-ms hamming window and a 10-ms frame shift. Each frame of the speech data is represented by a 39-dimensional feature vector that consists

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

of 13 MFCCs with their deltas and double-deltas. Then, the speaker-based mean and variance normalization using cepstral mean and variance normalization (CMVN) are applied to both the training and testing features. Seven consecutive feature frames are spliced to 40 dimensions via linear discriminative analysis (LDA) and maximum likelihood linear transformation (MLLT), which is a feature orthogonalizing transform, is applied to make the features more accurately modeled by diagonal covariance Gaussians. Moreover, to address speaker variability, speaker adaptive training (SAT) is performed using feature-space maximum likelihood linear regression (fMLLR). Using these features, we conducted experiments to determine the number of states and the HMM topologies for the phone and syllable acoustic modeling units. According to the results, the 3-state left-to-right HMM topology with the fourth-last non-emitting state is better for the phone acoustic modeling units. However, the syllable units are longer and subject to substantial acoustic variation, and a model of such units are expected to have many HMM states. Therefore, the 5-state left-to-right HMM topology with the sixth-last non-emitting states is suitable for the syllable acoustic modeling units.

In the Kaldi setup, the numbers of tied states and Gaussians depend on the number of hours of training speech to avoid the problems of overfitting and underfitting. We started from the Iban language example setup since its training hours are more similar to ours compared to the other examples. The number of leaves was 4200 and the total number of Gaussians was 40000. For all context-dependent acoustic units, after several tunings, we attain the best performances at 3948, 3532, and 3448 tied states with 50000 Gaussian mixtures for the syllable, rounded phone, and phone units, respectively. The numbers of context-independent states for the syllable, rounded phone, and phone units are 1165, 122, and 107, respectively, and all the context-independent acoustic units use the same numbers of Gaussian mixtures as the context-dependent acoustic units. As a result, for DNN training, the dimensions of the output layers are 3948, 3532, and 3448 senones for the CD syllable, CD rounded phone, and CD phone units, respectively, whereas for the context-independent models, the dimensions of the output layers are 1165, 122, and 107 for the syllable, rounded phone, and phone units, respectively.

A word-based backoff trigram language model is built using the SRI language modeling (SRILM) toolkit [49]. The model is trained using a text corpus of 216,076 sentences and smoothed using the modified Kneser-Ney smoothing algorithm. The perplexity value of the model is 76 on the 5k development test set sentences and it has an out of vocabulary (OOV) rate of 5%. This language model is applied for all acoustic modeling units: CD syllables, CD phones, CD rounded phones, CI syllables, CI phones, and CI rounded phones.

For our DNN-HMM experiments, we developed DNN models that train using single-task learning (STL) and MTL schemes. The network configuration for single-task DNN training is as follows: 40-dimensional higher resolution MFCC features and 100-dimensional i-vector speaker adaptation features are used as input features. A left context width of 9 and a right context width of 7 are used to combine the frames. The dimension of the input layer is 140. The pnorm nonlinearity [50] is used as an activation function, which is a dimension-reducing nonlinearity, and pnorm units with a group size of 8 and a p-value of two are used for all the acoustic modeling units. The greedy layer wise supervised pretraining algorithm is used for network initialization, and all the DNN models are trained using the preconditioning stochastic gradient descent optimization algorithm to minimize the cross-entropy training criteria with a mini-batch size of 512 frames. We use exponential learning rate scheduling, in which the initial learning rate is 0.005 for the first epoch and is decreased exponentially to the final learning rate of 0.0005 for the last epoch during training [50]. We use the Kaldi toolkit [48] as the DNN training tool. The training process is accelerated using the Nvidia GeForce GTX 1050 GPU. We use the standard values of the decoding parameters that are embedded in the toolkit, such as the decoding beam, lattice beam, acoustic and language model scales, and word insertion penalty. A weighted finite state transducer is used for decoding.

By making the above parameters consistent for all acoustic modeling units, we conduct several preliminary experiments to decide the optimal values of major hyperparameters, namely, the number of hidden layers, the size of hidden layer dimensions, and the number of epochs. For tuning the number of hidden layers, we used the pnorm hidden layers with input dimensions of 2400 and output dimensions of 300 and epochs with value of 10. The tuning results are presented in Figure 2. The results show that the optimal number of hidden layers is four for all acoustic modeling units. To tune the size of hidden layer dimensions, we use four hidden layers and 10 epochs. The input dimensions of 2000 and output dimensions of 250 are the optimal sizes of the pnorm hidden layer dimensions for all acoustic units as presented in Figure 3. The optimal number of epochs is eight for all acoustic modeling units which is achieved after tuning for several epochs by making the number hidden layers four and the size of the hidden layer input dimensions of 2000 and

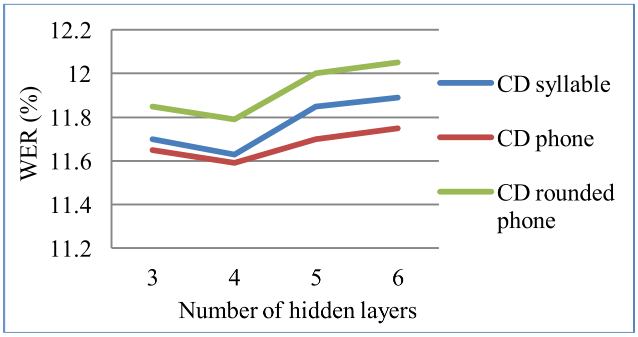

**FIGURE 2.** WER results vs. the number of hidden layers for all fundamental acoustic modeling units.

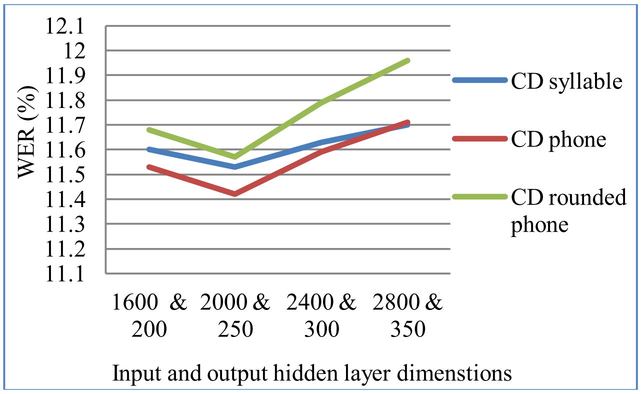

**FIGURE 3.** WER results vs. the size of hidden layer dimensions for all fundamental acoustic modeling units.

output dimensions of 250. The tuning results are presented in Figure 4.

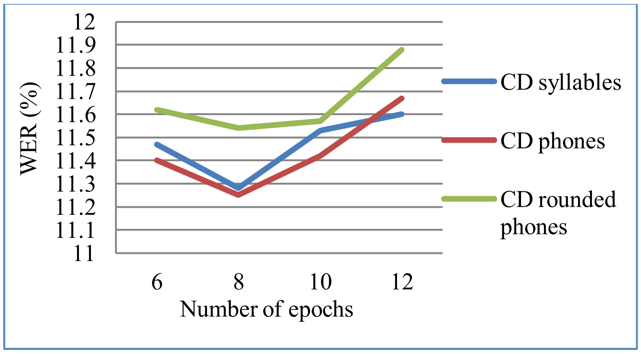

**FIGURE 4.** WER results vs. the number of epochs for all fundamental acoustic modeling units.

For multitask DNN training, we use all the model parameters of the single-task DNN except the task weight values. The task-weight values for two-task and three-task DNN training are tuned and the results are presented in Tables 8 and 9, respectively. According to the results, the task weight values of 0.7 and 0.3 for primary and ancillary tasks are optimal values for two-task learning while task weight values of 1, 0.7, and 0.3 for primary task and first and second auxiliary tasks, respectively, are optimal values for three-task learning.

Using the optimal values of the major hyperparameters which are obtained from the above preliminary experiments and other parameters which are described in this sub-section, we have conducted several actual experiments as described in sub-section B.

### B. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the different experiments with their results analysis and discussions.

#### 1) EXAMINE THE DEVELOPMENT OF FUNDAMENTAL-ACOUSTIC-UNIT-BASED STL-DNN MODELS

The first experiment explores the development of the fundamental-acoustic-unit-based DNN models for the Amharic language, as presented in Tables 5 and 6. According to the results of the experiments, all the basic-acoustic-modeling-unit-based DNN models outperform the baseline GMM models. In the CD unit experiment, as presented in Table 5, the CD phone and CD syllable outperform the CD rounded phone. The CD syllable realizes 11.28% WER, which corresponds to absolute and relative WER reductions of 2.36% and 17.3% over the equivalent GMM model. The CD phone realizes 11.25% WER, which represents 2.24% absolute and 16.6% relative performance improvement over the corresponding GMM model. The CD rounded phone outperforms the GMM model by absolute and relative WER reductions of 1.9% and 14.14%, respectively.

**TABLE 5.** WERs (%) of the baseline context-dependent models.

| Model | CD syllable | CD phone | CD rounded phone |
|---|---|---|---|
| GMM | 13.64 | 13.49 | 13.44 |
| STL-DNN | 11.28 | 11.25 | 11.54 |

**TABLE 6.** WERs (%) of the context-independent models.

| Model | CI syllable | CI phone | CI rounded phone |
|---|---|---|---|
| GMM | 16.61 | 21.75 | 20.56 |
| STL-DNN | 13.10 | 16.68 | 16.95 |

In the CI unit experiment, as listed in Table 6, the CI syllable outperforms the baseline syllable-based GMM model with 3.51% and 21.13% absolute and relative performance improvement, respectively. Similarly, the CI phone and CI rounded phone realize absolute performance enhancements of 5.07% and 3.61%. Hence, the relative performance improvements over the baseline models are 23.31% and 17.56%, respectively. As a result, in both experiments, the CD- and CI-acoustic-unit-based DNN models outperform the baseline conventional GMM models with an absolute WER reduction of 1.9 to 5.07%. These results demonstrate that the DNN acoustic modeling technique substantially improves the performance of LVCSR systems for the Amharic language.

Moreover, the rounded phone acoustic modeling units are investigated to analyze the influence of the labiovelars and labialized phones on the performance of the LVCSR system in comparison with the fundamental phone units. The experimental results that are presented in Tables 5 and 6 demonstrate that the rounded phones are outperformed by 2.5% and 1.6% by the basic phone units in the context-dependent and context-independent speaker-adapted models. Similarly, according to the speaker-independent experimental results, which are presented in Tables 10 and 11, the rounded phones are outperformed by 0.9% and 1.7% by the basic phones in the context-dependent and context-independent models, respectively. The experimental results demonstrate that considering the rounded nature of the rounded phones does not improve the performance of the DNN models. Therefore, the basic phones are more suitable than the rounded phones as acoustic modeling units for developing phone-based DNN models.

### 2) EFFECT OF SPARSELY DISTRIBUTED SYLLABLES ON THE PERFORMANCE OF LVCSR SYSTEM

In this experiment, we examine the effect of an uneven distribution of syllables in the training dataset on the performance of the Amharic syllable-based LVCSR system. The experiment is conducted by removing the rarely occurring syllables from the phone list and lexical dictionary. The syllables that have frequencies that are below 50 are removed, and 210 syllables remain from a total of 233 syllables of the language. Then, both the GMM and DNN acoustic models are trained using 210 syllables with the model parameters that are defined in the experimental setup in Section V-A. Table 7 presents the experimental results. Since these models are built using the evenly distributed syllables in the training data, the CD syllable attains a 0.48% absolute WER reduction using the GMM model compared with the GMM model result that is presented in Table 6, for which the model was built using all the syllables. Correspondingly, the DNN model that is built using the CD syllable obtains a 0.18% absolute performance improvement over the baseline results that are presented in Table 5. The performances of both the CD-syllable-based GMM and DNN models improve when they are trained using the evenly distributed syllables in the limited training corpus. However, the performance of the CD-syllable-based GMM model is better improved than the DNN model relative to the baseline GMM and DNN models, respectively. This finding is because the irregular distribution of syllables in the training corpus has a larger effect on the performance of the GMM model than on the DNN model since the DNN can model rarely occurred modeling units. Therefore, even if the DNN models reduce the data sparsity and unevenness problems to some extent, the performances of both the GMM and DNN models that are trained using the CD syllable acoustic units are influenced by data sparseness and unevenness.

**TABLE 7.** WERs of models that are based on evenly distributed syllables.

| Model | CD syllable (%) |
|---|---|
| GMM | 13.16 |
| DNN | 11.10 |

### 3) COMPARISON OF SPEAKER-INDEPENDENT AND SPEAKER-ADAPTED DNN MODELS

In this experiment, we compare the speaker-independent and speaker-adapted DNN models in terms of performance in reducing the data sparsity and uneven distribution of syllable units in the limited training dataset. Initially, we build the speaker-independent DNN models, which train over LDA and MLLT alignment, and we compare the obtained results with the experimental results that are presented in Tables 5 and 6. Therefore, for the speaker-independent model investigation, we used 3365, 2873, and 2866 tied states as DNN targets for syllable, rounded phone, and phone context-dependent units and 1165, 122, and 107 context-independent states as DNN targets for syllable, rounded phone, and phone context-independent units, respectively.

Tables 10 and 11 present the performances of the CD and CI DNN models that are trained over LDA and MLLT alignment, respectively. According to Table 10, the DNN models that are trained using the CD syllable, CD phone, and CD rounded phone units realize absolute WER reductions

**TABLE 8.** WER (%) of CD syllable-CI syllable acoustic units based two-task DNN model with different task weight values.

| Task weight values | | WER (%) |
|---|---|---|
| Primary task | Ancillary task | |
| 0.9 | 0.1 | 10.98 |
| 0.8 | 0.2 | 10.93 |
| 0.7 | 0.3 | 10.86 |
| 0.6 | 0.4 | 10.97 |
| 0.5 | 0.5 | 11.12 |

**TABLE 9.** WER (%) of CD syllable-CI phone-CI syllable acoustic units based three-task DNN model with different task weight values.

| Task weight values | | | WER (%) |
|---|---|---|---|
| Primary task | Auxiliary task-1 | Auxiliary task-2 | |
| 1 | 0.9 | 0.1 | 10.96 |
| 1 | 0.8 | 0.2 | 10.89 |
| 1 | 0.7 | 0.3 | 10.84 |
| 1 | 0.6 | 0.4 | 10.94 |
| 1 | 0.5 | 0.5 | 11.08 |

**TABLE 10.** WERs (%) of the speaker-independent and context-dependent GMM and DNN models.

| Model | CD syllable | CD phone | CD rounded phone |
|---|---|---|---|
| GMM | 17.22 | 15.68 | 15.85 |
| STL-DNN | 12.32 | 11.59 | 11.69 |

**TABLE 11.** WERs (%) of context-independent and speaker-independent GMM and DNN models.

| Model | CI syllable | CI phone | CI rounded phone |
|---|---|---|---|
| GMM | 19.63 | 24.60 | 24.28 |
| STL-DNN | 13.59 | 17.80 | 18.10 |

of 4.9%, 4.09%, and 4.16%, respectively, over the corresponding GMM models. Here, the CD syllable unit significantly reduces the relative WER by 2.36% and 2.21% over the equivalent CD phone and CD rounded phone units. Similarly, according to Table 11, the DNN models that are trained using the CI syllable, CI Phone, and CI rounded phone units realize 6.04%, 6.8%, and 6.18% absolute performance improvements, respectively, over the baseline GMM models. In addition, the CI syllable units realize significant performance improvements over the CI phone and CI rounded phone units, with relative performance improvements of 3.13% and 5.32%, respectively. Overall, the experimental results demonstrate that both CD and CI syllable units realize significant absolute WER reductions over the equivalent other units. Hence, the GMM model that is trained using the syllable units suffer from the problems of data sparsity and uneven distribution of syllables in the training data, while DNN reduces such limitations of the GMM and realizes larger performance improvement using the syllable units compared to the other units.

According to the experimental results that are presented in Tables 5 and 6, the ivector speaker-adapted syllable-based DNN model outperforms the equivalent GMM. However, the improvement is smaller than that of the speaker-independent DNN model that is trained over LDA and MLLT results, according to Tables 10 and 11. Comparing the two experimental results, the speaker-adapted and speaker-independent DNN models that are trained using CD syllable units realize absolute performance improvements of 2.36% and 4.9%, respectively. In contrast, the speaker-adapted and speaker-independent DNN models that are trained using the CI syllable units realize absolute performance improvements of 3.51% and 6.04%, respectively. Hence, the performances of the speaker-independent syllable-based DNN models have improved more than those of the speaker-adapted syllable-based DNN models over the baseline GMM models. Thus, the data sparsity and uneven distribution of the syllable units are better mitigated using the speaker-independent DNN models compared to the speaker-adapted DNN models. However, the speaker-adapted syllable-based DNN models consider the speaker variety; hence, they realize higher overall accuracy than the speaker-independent syllable-based DNN models. Both the speaker-independent and speaker-adapted syllable-based DNN models have reduced the sparsity and unevenness of the syllable units in the limited training dataset. However, those models still suffer from the lack of a sufficient training dataset. Section V-B-4 suggests possible solutions for overcoming these problems.

### 4) INVESTIGATION OF HYBRID ACOUSTIC MODELING UNITS USING THE MTL-DNN PARADIGM

According to the experimental results that are presented in Sections V-B 2 and 3, the performances of the CD-syllable-unit-based GMM and DNN models are affected by the sparsity and unevenness of the syllables in the training dataset, in addition to the scarcity of sufficient training corpora. To overcome these problems, this experiment proposes different hybrid acoustic modeling units by training the fundamental acoustic units via the MTL-DNN paradigm, as presented in Table 12 and Figure 5. These units are investigated by dividing them into four major categories: In the first category, the context-dependent acoustic units are used as the primary task to jointly train with the corresponding context-independent acoustic units as an auxiliary task. CD syllable with CI syllable and CD phone with CI phone are the proposed hybrid units that are included in this category, as presented in the first two rows of Table 12.

**TABLE 12. Performances of various hybrid-acoustic-unit-based MTL-DNN models.**

| Remark | Main Task | Auxiliary tasks | WER (%) |
|---|---|---|---|
| Proposed units | CD phone | CI phone | 11.01 |
| | CD syllable | CI syllable | 10.86 |
| | CD syllable | CD phone | 11.13 |
| | CD syllable | CD syllable | **10.81** |
| | CD syllable | CD phone and CI syllable | 10.84 |
| | CD syllable | CD syllable and CI syllable | **10.79** |

In the second category, the context-dependent units are used as the primary task to train together with alternative context-dependent units, which are trained using various tied states as an auxiliary task. This category includes the CD syllable along with another CD syllable that is trained using a different number of senones as shown in the fourth row of Table 12, and Table 13. In the third category, the context-dependent units are used as the primary task to mutually train with other context-dependent units as an auxiliary task, which includes the CD syllable with the CD phone, as presented in the third row of Table 12. All the proposed hybrid acoustic units that are in the first, second, and third categories are trained using the two-task learning paradigm. In addition, there is a fourth category that combines the second- and third-category hybrid acoustic units with the CI syllable for training additional hybrid acoustic units. This category includes the CD syllable-CD phone with the CI syllable and the CD syllable-CD syllable with the CI syllable. These units are trained together using the three-task learning approach, as presented in the last two rows of Table 12.

**TABLE 13. WERs of the MTL-DNN models in which the CD syllable is jointly trained with an alternative CD syllable using various numbers of senones.**

| Number of Senones | 700 | 800 | 900 | 1000 | 1300 |
|---|---|---|---|---|---|
| WER (%) | 10.81 | 10.89 | 10.89 | 10.98 | 11.01 |

In the first category, we examine the joint training of the CD syllable as the main task with the CI syllable as an auxiliary task. This approach realizes the best WER of 10.86%,

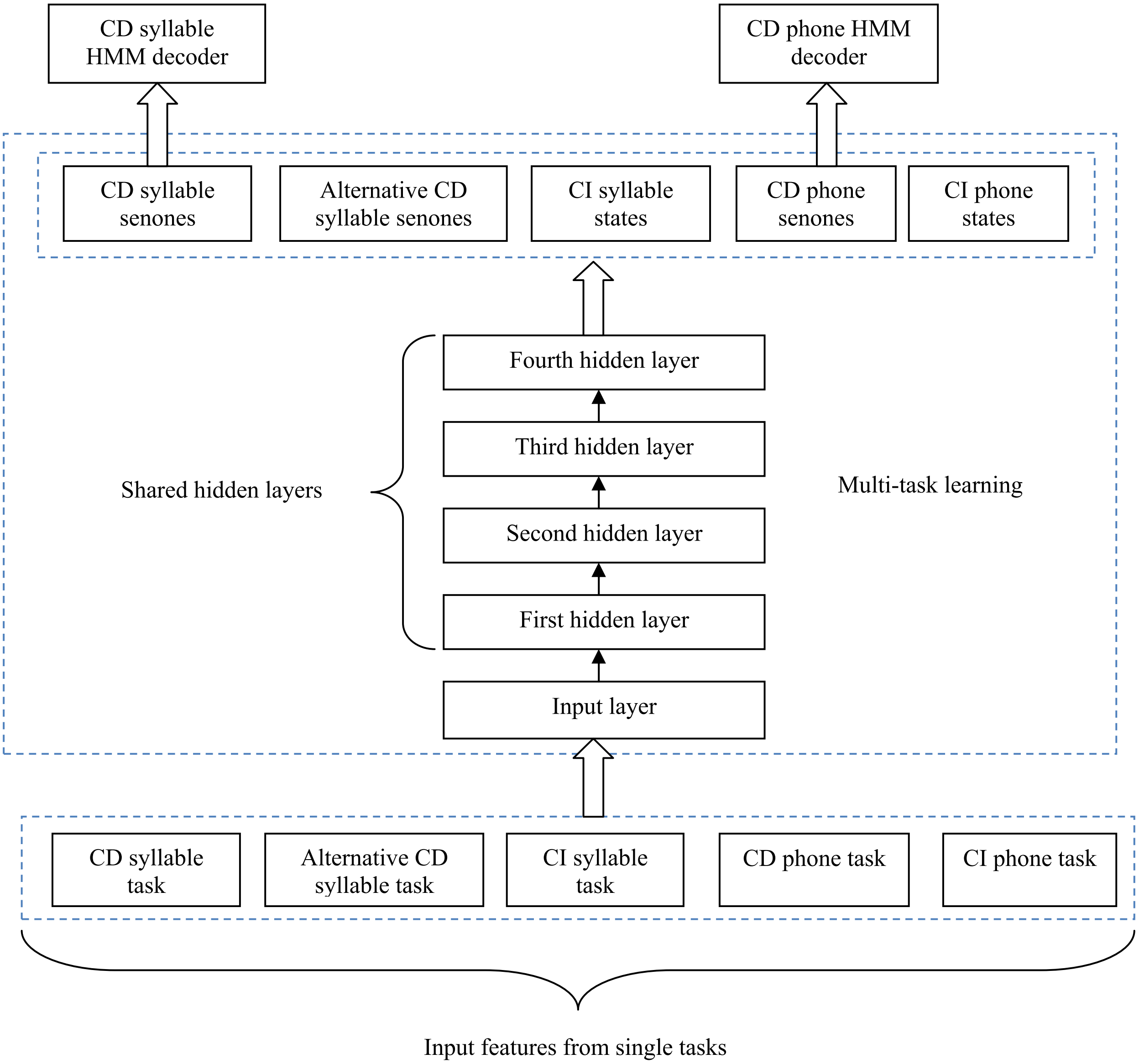

**FIGURE 5.** Architectural representation of the proposed hybrid acoustic modeling units.

which represents a 4.14% relative performance improvement over the corresponding CD syllable STL-DNN model. Similarly, we investigate the joint training of the CD phone as the main task with the CI phone as an auxiliary task; it realizes 11.01% WER, which represents a 2.13% relative performance gain over the baseline CD phone STL-DNN model. However, this result proves that the CI phone unit does not substantially facilitate the performance of the CD phone unit. This finding is because the CD-phone-based model does not suffer from data sparsity or unbalanced distribution of phones in the training dataset. According to the result, the CD-syllable-based hybrid unit outperforms the CD-phone-based hybrid unit. Thus, we use the CD syllable as the primary task and other units as auxiliary tasks throughout the remaining multi-task learning experiments, as presented in Table 12.

In the second category, we analyze the joint training of the CD syllable as the primary task with the CD syllable that is trained using a different number of senones as an auxiliary task, as presented in Table 13. The results demonstrate that the best performance is realized from the training of the CD syllable with another CD syllable that is trained using 700 senones as an ancillary task, namely, 10.81% WER. This result represents an absolute performance improvement over the baseline CD syllable STL-DNN model of 0.47%. In addition, Table 13 shows the influence of auxiliary task on the performance of the primary task as a function of the number of senones on which the auxiliary task is trained. As the number of senones on which the ancillary task is trained is increased and approaches the number of senones of the primary task, the primary task obtains less assistance from the auxiliary task and the performance of the corresponding hybrid model is reduced. However, the performance of the primary task is improved when the number of senones of the auxiliary tasks is 4-7 times less than the number of senones of the primary task. Moreover, the CD phone is not assisted substantially by the ancillary task for the CD phone with the CI phone hybrid units in category one. Therefore, we did not explore the effect of training the CD phone with the various numbers of senones as the auxiliary task together with the corresponding CD phone as the primary task in the hybrid units.

In the third category, the CD syllable as the primary task with the CD phone as an auxiliary task are trained together and realize a 0.28% absolute WER reduction over the baseline single-task CD syllable unit. Similarly, in the fourth category, the CD syllable as the primary task, and the CD phone and the CI syllable train together as auxiliary tasks and realize an absolute WER reduction of 0.44%. Another hybrid unit that combines the CD syllable as the primary task with the CD syllable trains using 700 tied states and the CI syllable as auxiliary tasks and realizes the lowest WER of 10.79%, which represents 0.49% absolute and 4.34% relative WER reductions over the baseline CD-syllable-based single-task DNN model.

### 5) RECOGNITION SPEEDS OF HYBRID AND FUNDAMENTAL ACOUSTIC MODELING UNITS

Table 14 presents the decoding speeds of the hybrid and CD fundamental acoustic units. Of the hybrid acoustic units, the CD syllable-CD syllable-CI syllable model realizes the fastest decoding speed with a real-time factor (RTF) of 1.02069, while the CD syllable-CD phone-CI syllable hybrid-unit-based model exhibits the slowest decoding speed with a real-time factor of 1.31444. This finding is because this model uses a larger decoding beam size than the universal decoding beam size, as specified in Section V-A. The CD syllable-CD syllable and the CD syllable-CD syllable-CI syllable units have realized the highest recognition accuracy, as stated in Section V-B-4. Thus, both units outperform other hybrid units in terms of the recognition speed and accuracy. In contrast, the fundamental units, namely, CD rounded phone, CD phone, and CD syllable units, realizes the fastest, faster and fast decoding speeds, respectively, and the decoding speed is inversely proportional to the recognition accuracy, which is reported in Section V-B-1. Moreover, Table 15 lists the decoding speeds of the context-independent units. The CI syllable unit is faster than the CI phone and rounded phone units with a real-time factor of 1.02176.

**TABLE 14.** Decoding speeds of the various context-dependent and hybrid acoustic modeling units.

| Remark | Acoustic Unit | RTF |
|---|---|---|
| Baseline units | CD syllable | 1.02178 |
| | CD rounded phone | **1.01928** |
| | CD phone | 1.02049 |
| Proposed units | CD syllable-CI syllable | 1.02255 |
| | CD syllable-CD syllable | **1.02079** |
| | CD syllable-CD phone | 1.02131 |
| | CD syllable-CD phone-CI syllable | 1.31444 |
| | CD syllable-CD syllable-CI syllable | **1.02069** |

**TABLE 15.** Decoding speeds of the context-independent acoustic modeling units.

| Acoustic Unit | RTF |
|---|---|
| CI syllable | **1.02176** |
| CI rounded phone | 1.04560 |
| CI phone | 1.04678 |

### 6) EFFECT OF THE TRAINING DATASET SIZE ON THE PERFORMANCES OF STL AND MTL DNN MODELS

To analyze the influence of the size of the dataset on the performances of the fundamental- and hybrid-acoustic-unit-based STL and MTL-DNN models, we have prepared two training of 7 hours and 13 hours in length via random selection from 26-hour training dataset. Prior to training the hybrid-unit-based MTL-DNN models, the fundamental-unit-based STL-DNN models are developed. Those models are trained with the parameters that are specified in the experimental setup in Section V-A, except for the tied states and the hidden layer dimensions. Since the training dataset is relatively small, the average numbers of tied states that are used are 2156 and 1706 for the CD syllable and CD phone acoustic units, respectively. Moreover, the hidden-layer input and output dimensions are reduced from the baseline values of 2000 and 250 nodes to 1000 and 125 nodes, respectively, to overcome model overfitting.

Table 16 presents the performances of the fundamental-acoustic-unit-based STL-DNN models that were trained with datasets of various sizes. Using the 7-hour dataset, the CD-phone-based STL-DNN model outperforms the CD-syllable-based STL-DNN model with an absolute WER reduction of 5.02%. Similarly, the CD-phone-based model also outperforms the CD-syllable-based model with an absolute WER reduction of 2.24% when the models are trained using the

**TABLE 16.** WERs (%) of the baseline CD- and CI-unit-based STL-DNN models.

| Acoustic Unit | 7 Hrs | 13 Hrs | 26 Hrs | Best case absolute WER (%) reduction |
|---|---|---|---|---|
| CD syllable | 25.11 | 18.29 | 11.28 | **13.85** |
| CD phone | **20.09** | **16.05** | **11.25** | 8.84 |
| CI syllable | 28.91 | **22.33** | **13.10** | **15.81** |
| CI phone | **26.84** | 23.06 | 16.68 | 10.16 |

13-hour dataset. Hence, the performance of the CD-syllable-based STL-DNN model is improved when the training dataset size is substantially increased compared to the equivalent CD-phone-based models. Hence, the CD-syllable-based models realized a larger best-case absolute WER reduction of 13.85% compared to the CD-phone-based models.

Table 17 presents the performance of the proposed hybrid modeling units as a function of the training dataset size. The results demonstrate that all the investigated hybrid-acoustic-unit-based MTL-DNN models outperform the equivalent single-acoustic-unit-based models. Using the 7- and 13-hour training datasets, the CD phone-CI phone-based model outperforms the CD-syllable-based hybrid units. This result is because there are many sparse and unevenly distributed syllables in both limited datasets. For instance, there are 13 syllables that are not found in the 7-hour training dataset. However, using the 26-hour training dataset, the CD syllable units that are jointly trained with other units as a hybrid unit outperform the CD phone-CI phone hybrid unit and realize larger best-case absolute WER reductions of 12.91 to 13.69%.

**TABLE 17. WERs (%) of the proposed hybrid unit-based MTL-DNN models.**

| Main Task | Auxiliary tasks | 7 Hrs | 13 Hrs | 26 Hrs | Best case absolute WER (%) reduction |
|---|---|---|---|---|---|
| CD phone | CI phone | **19.14** | **15.29** | 11.01 | 8.13 |
| CD syllable | CI syllable | 24.11 | 15.93 | 10.86 | 13.25 |
| CD syllable | CD phone | 24.16 | 16.05 | 11.13 | 13.03 |
| CD syllable | CD syllable | 24.50 | 16.22 | **10.81** | **13.69** |
| CD syllable | CD phone and CI syllable | 23.94 | 16.22 | 10.84 | 13.10 |
| CD syllable | CD syllable and CI syllable | **23.70** | **16.03** | **10.79** | 12.91 |

### 7) COMPARISON OF HYBRID AND FUNDAMENTAL ACOUSTIC MODELING UNITS

The hybrid acoustic modeling units that are proposed in Section V-B-4 are compared with the fundamental acoustic modeling units, namely, the CD syllable and the CD phone, in terms of recognition performance and speed. All the hybrid modeling units outperform the baseline CD syllable and CD phone units with relative performance improvements of 1.33 to 4.34%. This finding is because when the fundamental acoustic modeling units are jointly trained via MTL, the training data size is increased and, hence, the risk of overfitting and the Rademacher complexity of the model are reduced. As a result, the data scarcity and sparsity challenges are mitigated, and the performances of the hybrid-acoustic-unit-based models are improved.

The hybrid units, namely, CD syllable-CI syllable, CD syllable-CD syllable, and CD syllable-CD syllable-CI syllable, outperform the equivalent hybrid units, such as CD syllable-CD phone and CD syllable-CD phone-CI syllable. This result is because the hybrid units in the first set share the same labeled unit training dataset, while the hybrid units in the second set use different labeled unit training datasets. The higher performing hybrid acoustic units are CD syllable-CD syllable-CI syllable and CD syllable-CD syllable, which realize WERs of 10.79% and 10.81%, respectively. The results demonstrate that the CD syllable-CD syllable-CI syllable and CD syllable-CD syllable units outperform the corresponding single-task CD syllable unit with 4.34% and 4.17% relative performance improvements, respectively. Consequently, the highest performing hybrid acoustic modeling unit is the CD syllable-CD syllable-CI syllable unit.

In contrast, the CD syllable-CD phone unit exhibits the lowest performance of the hybrid acoustic modeling units. The CD phone-CI phone hybrid unit outperforms the singly training CD phone unit with a relative performance improvement of 2.13%. However, the performance improvement is small compared to the other hybrid acoustic modeling units. Moreover, the performance of the hybrid acoustic unit that combines the CD syllable as the primary task with the CD syllable that is trained using a different number of senones as an auxiliary task depends highly on the number of senones in the auxiliary task. Thus, it is advantageous to use fewer senones for an ancillary task than the number of senones of the primary task.

Using training datasets of various sizes, all the proposed hybrid acoustic modeling units consistently outperform the equivalent basic acoustic units, namely, the syllable and phone units, as shown in Figure 6. Typically, as the training data size increases, the performances of the hybrid-unit-based DNN models improve and higher performance is realized from the joint training of the CD syllable with other units. Hence, the proposed hybrid acoustic modeling units can improve the performance of LVCSR of the Amharic language under both limited and sufficient training dataset conditions. Moreover, the performances of both the syllable- and phone-acoustic-unit-based DNN models improve when the training dataset increases in size, as discussed in Section V-B-6. The syllable-based model substantially outperforms the phone-based model. The CD-phone-based models outperformed the syllable-based model using limited training datasets (7 and 13 hours). Therefore, if the available training data corpus is sufficiently large and only the accuracy is prioritized, the single CD syllable unit is the better performing acoustic modeling unit and to improve the performance, the CD syllable should be trained jointly with other units as a hybrid unit using the MTL technique. Alternatively, if the training dataset is limited and both the accuracy and the decoding speed of the recognizer are prioritized, the CD phone modeling unit is the better performing acoustic unit and to improve its performance further, the CD phone-CI phone hybrid acoustic modeling units should be trained via the MTL paradigm.

Based on the recognition speed, the use of secondary tasks increases the computational cost of training compared to the single task, although the difference in recognition speed is

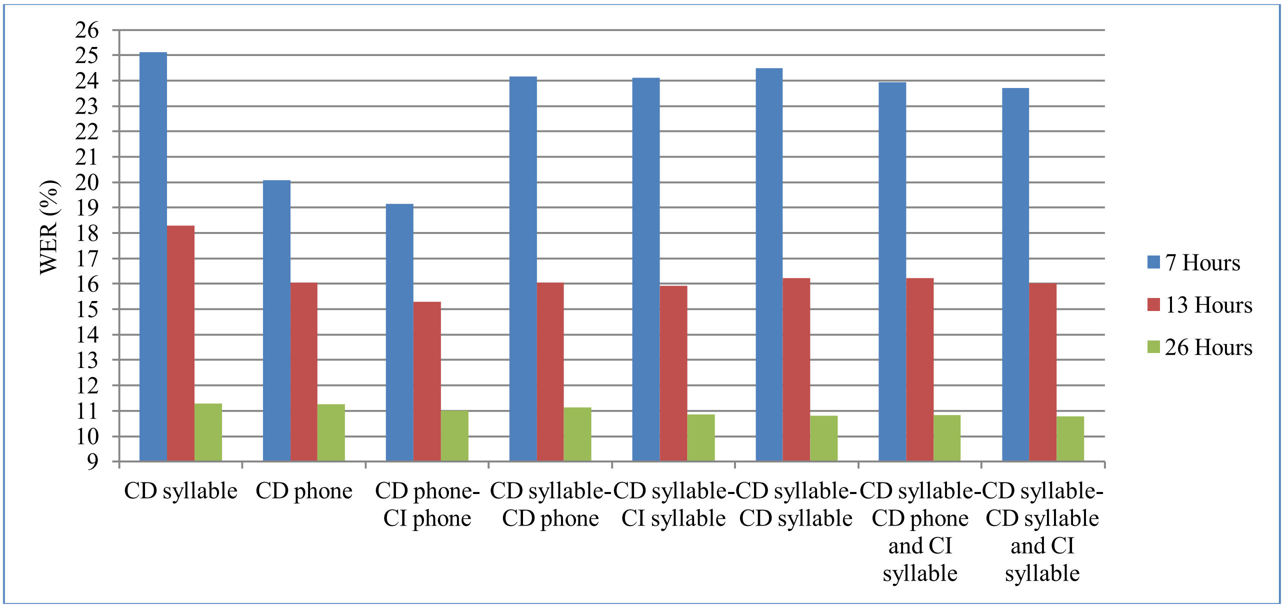

**FIGURE 6.** Comparison of hybrid and fundamental acoustic modeling units.

not substantial since the secondary tasks are discarded at the end of the training process. Thus, the decoding speed of a hybrid-acoustic-unit-based MTL model that is trained using the CD syllable as a primary task with various secondary tasks is almost the same as that of the CD-syllable-based single-task DNN model. In line with this, the CD-rounded-phone-based and CD-phone-based STL-DNN models have faster decoding speeds than the basic-CD-syllable-based STL-DNN and hybrid-CD-syllable-unit-based MTL-DNN models. This finding is because the decoding speed is varied due to the graph size differences among the acoustic modeling units, thereby rendering all the decoding parameters that are specified in Section V-A universal for all models. The graph sizes of all the hybrid-CD-syllable- and basic-CD-syllable-unit-based models are larger than those of the equivalent CD-phone- and CD-rounded-phone-unit-based models.

Moreover, the fundamental CI acoustic modeling units, namely, the CI syllable and CI phone acoustic units, are compared in terms of performance and decoding speed. The CI syllable significantly outperforms the corresponding CI phone with an absolute performance improvement of 3.58%. This result is because CI syllable units are long and stronger in terms of the co-articulation effect of the neighboring syllable than the phone units. The CI syllable unit also has the faster decoding speed than the corresponding CI phone units. Hence, the CI CV syllable is the most suitable acoustic modeling unit for building the CI DNN model for Amharic speech recognition.

## VI. CONCLUSIONS AND FUTURE WORK

This study proposed hybrid acoustic modeling units by jointly training the fundamental acoustic units via the MTL-DNN modeling technique for LVCSR of a low-resource language, namely, Amharic. First, this study developed the single-task DNN models using the fundamental acoustic modeling units, namely, phone, rounded phone, and syllable units. The experimental results demonstrate that all the single-task DNN models outperform the corresponding baseline conventional GMM-HMM models, with absolute WER reductions of 1.9 to 5.07%. However, the fundamental acoustic-unit-based single-task DNN models are suffered from the problems of training data scarcity, sparsity, and unevenness, especially CD syllable unit-based models. Second, to overcome the above problems, this study proposed various hybrid acoustic modeling units through the joint training of the fundamental acoustic modeling units using the MTL-DNN paradigm. Thus, the experimental results demonstrate that all the proposed hybrid-acoustic-unit-based DNN models outperform the singly trained fundamental-acoustic-unit-based DNN models with relative performance improvements of 1.33 to 4.34%. The best performing hybrid acoustic units is the CD syllable-CD syllable-CI syllable unit with a WER of 10.79%. Third, the recognition performances of the fundamental and hybrid acoustic modeling units are compared using training datasets of various sizes. The experimental results indicate that the performances of all the fundamental and hybrid acoustic units improve as the training dataset size increases. Especially, the performances of the CD-syllable-based models improve significantly. Thus, all the hybrid-acoustic-units-based DNN models outperform the basic-acoustic-units-based models via different training dataset sizes with the WER reductions of 2.13 to 5.62%. Hence, the CD syllable acoustic modeling units perform better if and only if a sufficient training corpus is available and the accuracy of the recognizer is prioritized. The CD phone units are the superior choice for acoustic modeling units if the available training dataset is limited and the accuracy and speed of the recognizer are prioritized. The hybrid acoustic modeling units are the best performing acoustic modeling units if the accuracy of recognizer is prioritized under both limited and sufficient training dataset conditions.

The findings of this study will facilitate the work of language researchers and that of companies that are developing speech recognition applications for this language. Moreover, we recommend the investigation of various hybrid modeling units by jointly training different language specific fundamental acoustic modeling units via multitask learning for low-resource languages, which have limited training corpora with sparse and unevenly distributed acoustic modeling units. On the other hand, further investigation will focus on three issues: the expansion of training corpus, the acoustic modeling techniques, and the properties of the language. First, increase the size of training corpus can boost the performance of ASR system. Thus, we will explore the performances of hybrid acoustic units based DNN-HMM models by augmenting the available training corpus (particularly, the highly sparse syllables such as vu, vwa, kwi, žə, žwa, p'ə, p'u, p'i, p'e, and p'wa) using different audio augmentation techniques and by borrowing the training corpus from acoustically related resource-rich languages. Second, the acoustic modeling techniques such as the convolutional neural network (CNN), time delay neural network (TDNN), and long-short memory term memory (LSTM) deep neural network architectures, significantly outperform the ordinary DNN model [20], [51], [52]. Moreover, the sequential discriminative training criteria have higher discriminative power, which enhances the performance of the DNN ASR models related to those that are trained using the default cross-entropy criterion [53]. Therefore, we will explore the performances of hybrid acoustic modeling units that are trained via the CNN, TDNN, and LSTM acoustic modeling techniques with the discriminative training criteria. Third, Amharic is a morphologically rich language, in which the ASR models suffer from the OOV problem. This problem can be mitigated by using the smallest language and lexical modeling units other than the default word unit, namely, morphs [37]. We will investigate the impact of using morph units for language and lexical models with the hybrid acoustic units on the performance of the Amharic speech recognition systems.

## REFERENCES

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012. doi: 10.1109/TASL.2011.2134090.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012. doi: 10.1109/MSP.2012.2205597.

[3] X. Li, C. Hong, Y. Yang, and X. Wu, "Deep neural networks for syllable based acoustic modeling in Chinese speech recognition," in *Proc. APSIPA-ASC*, Kaohsiung, Taiwan, Oct./Nov. 2013, pp. 1–4.

[4] X. Li, Y. Yang, Z. Pang, and X. Wu, "A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition," *Neurocomputing*, vol. 170, pp. 251–256, Dec. 2015. doi: 10.1016/j.neucom.2014.07.087.

[5] A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky, and A. Y. Ng "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017. doi: 10.1016/j.csl.2016.06.007.

[6] L. Mateju, P. Cerva, and J. Zdansky, "Investigation into the use of deep neural networks for LVCSR of Czech," in *Proc. ECMSM*, Liberec, Czech Republic, Jun. 2015, pp. 1–4.

[7] H. Seki, K. Yamamoto, and S. Nakagawa, "Comparison of syllable-based and phoneme-based DNN-HMM in Japanese speech recognition," in *Proc. ICAICTA*, Bandung, Indonesia, Aug. 2014, pp. 249–254.

[8] M. Ahmed, P. C. Shill, K. Islam, A. Md Mollah, and M. A. H. Akhand, "Acoustic modeling using Deep Belief Network for Bangle Speech Recognition," in *Proc. 18th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2015, pp. 306–311.

[9] M. Cai, Y. Shi, and J. Liu, "Stochastic pooling maxout networks for low-resource speech recognition," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3290–3294.

[10] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 5592–5596.

[11] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 7, pp. 1172–1183, Jul. 2015.

[12] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. J. F. Gales, K. M. Knill, A. Ragni, H. Wang, and P. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *Proc. IEEE ASRU*, Scottsdale, AZ, USA, Dec. 2015, pp. 259–266.

[13] A. Dey, W. Zhang, and P. Fung, "Acoustic modeling for hindi speech recognition in low-resource settings," in *Proc. ICALIP*, Shanghai, China, Jul. 2014, pp. 891–894.

[14] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, Vancouver, BC, Canada, May 2013, pp. 7304–7308.

[15] R. Sahraeian, D. Van Compernolle, and F. de Wet, "Using generalized maxout networks and phoneme mapping for low resource ASR- a case study on flemish-afrikaans," in *Proc. (PRASA-RobMech)*, Port Elizabeth, South Africa, Nov. 2015, pp. 112–117.

[16] R. Sriranjani, B. M. Karthick, and S. Umesh, "Investigation of different acoustic modeling techniques for low resource Indian language data," in *Proc. 20th Nat. Conf. Commun. (NCC)*, Mumbai, India, Feb./Mar. 2015, pp. 1–5.

[17] W. Wang, W. Xu, X. Sui, L. Wang, and X. Liu, "Investigations of Low Resource Multi-Accent Mandarin Speech Recognition," in *Proc. IEEE Int. Conf. Inf. Automat.*, Lijiang, China, Aug. 2015, pp. 62–66.

[18] A. AbdAlmisreb, A. F. Abidin, and N. M. Tahir, "Maxout based deep neural networks for Arabic phonemes recognition," in *Proc. IEEE 11th Int. Colloq. Signal Process. Appl. (CSPA)*, Kuala Lumpur, Malaysia, Mar. 2015, pp. 192–197.

[19] M. Cai, Y. Shi, and J. Liu, "Deep maxout neural networks for speech recognition," in *Proc. IEEE ASRU*, Olomouc, Czech Republic, Dec. 2013, pp. 291–296.

[20] Y. Miao, F. Metze, and S. Rawat, "Deep Maxout Networks for low-resource speech recognition," in *Proc. IEEE ASRU*, Olomouc, Czech Republic, Dec. 2013, pp. 398–403.

[21] M. J. Gales, K. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *Proc. Int. Workshop SLTU*, May 2014, pp. 16–23.

[22] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 7689–7693.

[23] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP*, Vancouver, BC, Canada, May 2013, pp. 7319–7323.

[24] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, Vancouver, BC, Canada, May 2013, pp. 8619–8623.

[25] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *Proc. ICASSP*, Brisbane, QLD, Australia, Apr. 2015, pp. 4290–4294.
[26] P. Bell, P. L. Swietojanski, and S. Renals, P. Bell, P. Swietojanski, and S. Renals, "Multitask learning of context-dependent targets in deep neural network acoustic models," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 25, no. 2, pp. 238–247, Feb. 2017. doi: 10.1109/TASLP.2016.2630305.
[27] D. Chen and B. K. Mak, "Distinct triphone acoustic modeling using deep neural networks," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 2645–2649.
[28] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*, May 2013, pp. 6965–6969.
[29] M. Y. Tachbelie, S. T. Abate, L. Besacier, and S. Rossato, "Syllable-based and Hybrid acoustic models for Amharic speech recognition," in *Proc. 3rd Workshop SLTU*, Cape Town, South Africa, May 2012, pp. 5–10.
[30] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1601–1604.
[31] S. Ruder, "An overview of multi-task learning in deep neural networks," Jun. 2017, *arXiv:1706.05098*. [Online]. Available: https://arxiv.org/abs/1706.05098
[32] *Population and Housing Census 2007 Report, National*, Central Statistical Agency, Addis Ababa, Ethiopia, 2010.
[33] L. M. Lewis, G. F. Simons, and C. D. Fennig, "Amharic," in *Ethnologue: Languages of the World*, 18th ed. Dallas, TX, USA: SIL International, 2015.
[34] S., H/Mariam, S. P. Kishore, A. W. Black, R. Kumar, and R. Sangal, "Unit selection voice for Amharic using festvox," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, USA, Jun. 2004, pp. 103–107.
[35] Y. B. Gebremedhin, F. Duckhorn, R. Hoffmann, and I. Kraljevski, "A new approach to develop a syllable based, continuous amharic speech recognizer," in *Proc. Eurocon*, Zagreb, Croatia, Jul. 2013, pp. 1684–1689.
[36] S. T. Abate and W. Menzel, "Automatic speech recognition for an under-resourced language— Amharic," in *Proc. Interspeech*, Aug. 2007, pp. 1541–1544.
[37] M. Y. Tachbelie, S. T. Abate, and L. Besacier, "Using different acoustic, lexical and language modeling units for ASR of an under-resourced language—Amharic," *Speech Commun.*, vol. 56, pp. 181–194, Jan. 2014. doi: 10.1016/j.specom.2013.01.008.
[38] K. Tadesse, "Sub-word based amharic speech recognizer: An experiment using hidden Markov model (HMM)," M.S. thesis, Sch. Info. Stud. Afr., AAU, Addis Ababa, Ethiopia, 2002.
[39] S. T. Abate and W. Menzel, Wolfgang, "Syllable-based speech recognition for Amharic," in *Proc. SEMITIC-ACL*, Prague, Check Republic, Jun. 2007, pp. 33–40.
[40] M. M. Woldeyohannis, L. Besacier, and M. Meshesha, "Amharic speech recognition for speech translation," in *Proc. JEP-TALN-RECITAL*, Paris, France, Jul. 2016, pp. 114–124.
[41] A. E. Dribssa and M. Y. Tachbelie, "Investigating the use of syllable acoustic units for Amharic speech recognition," in *Proc. AFRICON*, Addis Ababa, Ethiopia, Sep. 2015, pp. 1–5.
[42] H. Seid and B. Gambäck, "A speaker independent continuous speech recognizer for Amharic," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 3349–3352.
[43] M. Y. Tachbelie and S. T. Abate, "Effect of language resources on automatic speech recognition for Amharic," in *Proc. AFRICON*, Addis Ababa, Ethiopia, Sep. 2015, pp. 1–5.
[44] (2013). *ELRA Catalogue*. [Online]. Available: http://catalog.elra.info
[45] D. Appleyard, "The Amharic language," in *Colloquial Amharic: Complete Course for Beginners*, 2nd ed. Abingdon, U.K.: Routledge, 2015, pp. 2–18.
[46] W. Leslau, "Phonology," in *Introductory Grammar Amharic*, Wiesbaden, Germany: Harrassowits Verlag, 2000, pp. 1–16.
[47] A. Teferra and G. Hudson, "Amharic sounds," in *Essentials Amharic*, vol. 18, 1st ed. Koln, Germany: Rüdiger Koppe, 2008, pp. 29–37.
[48] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, Waikoloa, HI, USA, 2011, pp. 1–4.
[49] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, CA, USA, Sep. 2002, pp. 901–904.
[50] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 215–219.
[51] M. Cai and J. Liu, "Maxout neurons for deep convolutional and LSTM neural networks in speech recognition," *Speech Commun.*, vol. 77, pp. 53–64, Mar. 2016.
[52] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.* , vol. 22, no. 10, pp. 1533–1544, Oct. 2014. doi: 10.1109/TASLP.2014.2339736.
[53] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, Mar. 2005, vol. 1, pp. 961–964.

**TESSFU GETEYE FANTAYE** received the M.Sc. degree in computer science from Addis Ababa University, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His current research interests include speech recognition, deep learning, and multimedia.

**JUNQING YU** received the Ph.D. degree in computer science from Wuhan University, in 2002. He is currently a Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His current research interests include digital video processing and retrieval, multimedia, and parallel algorithm.

**TULU TILAHUN HAILU** received the M.Sc. degree in computer science from Addis Ababa University, in 2013. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His current research interests include machine translation, deep learning, and multimedia.