An Amharic Syllable-Based Speech Corpus for Continuous Speech Recognition

Nirayo Hailu Gebreegziabher(✉) and Andreas Nürnberger

Fakultät für Informatik, Data and Knowledge Engineering Group, Otto von Guericke Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany
{nirayo.hailugebreegziabher, andreas.nuernberger}@ovgu.de

Abstract. Speech recognition systems play an important role in solving problems such as spoken content retrieval. Thus, we are interested in the task of speech recognition for low-resource languages, such as Amharic. The main challenges in solving Amharic speech recognition are the limited availability of corpora and complex morphological nature of the language. This paper presents a new corpus for the low-resource Amharic language which is suitable for training and evaluation of speech recognition systems. The corpus prepared contains 90 h of speech data with word and syllable-based annotation. Moreover, the use of syllable units for acoustic and language model in comparison with a morpheme-based model is presented. Syllable-based triphone speech recognition system provides a lower word error rate of 16.82% on the subset of the dataset. Moreover, syllable-based hybrid deep neural network with hidden Markov model provides a 14.36% word error rate.

Keywords: Speech recognition · Corpus · Neural and hidden Markov model · Syllable units

1 Introduction

With the increasing amount of spoken data being stored, shared, and processed nowadays, there is a need for systems performing automatic speech recognition, audio indexing, and search on audio streams. Hence, researchers are interested in the task of speech recognition and retrieving data from spoken contents, such as for Amharic. The domain of spoken contents includes broadcast news, oral historic archives, online lectures, meeting dialogues, and call-center conversations [14]. There are numerous amount of research that has been done on speech recognition [6, 7, 19–21]. However, performing speech recognition on low-resource languages raises some of the major research challenges in the area. There should be an open research with publicly available datasets and methodologies to speed up the progress in the field and to make speech recognition systems available for wider use.

There are efforts made to develop both morpheme-based [8, 9, 11] and syllable-based [7, 9] speech recognition systems for Amharic. However, all published works used only 20 h of training data [10]. In this paper, an effort has been made to collect

C. Martín-Vide et al. (Eds.): SLSP 2019, LNAI 11816, pp. 177–187, 2019.
https://doi.org/10.1007/978-3-030-31372-2_15

more Amharic speech and text corpora to make it publicly available for researchers in the field. We have also demonstrated the advantage of using Amharic syllable units instead of other units like morpheme.

The remainder of the paper is organized as follows: Sect. 2 describes the Amharic language. Section 3 discusses the corpus preparation. In Sect. 4, Amharic speech recognition components acoustic, language, and pronunciation models are described. Section 5 presents the experiments on the corpus. The last section, Sect. 6, provides discussion, conclusions, and highlights of the future work.

2 The Amharic Language

Amharic is the official language spoken in Ethiopia. It is a Semitic language of the Afro Asiatic Language group that is related to Hebrew and Arabic. There are more than 25 million users according to Ethnologue[1]. The language has its own writing system. As it is true in other languages, Amharic has its own phonetic and phonological characteristics. Amharic orthography, also known as ፊደል (fidəl), represents a consonant-vowel sequence, which is modified for the vowel.

There are seven vowels in Amharic namely, ኧ[ə], ኡ[u], ኢ[i], ኣ[a], ኤ[e], እ[ɨ], ኦ[o] [4, 5] (see Table 1). The language has 33 basic characters with each having seven forms for each consonant-vowel combination (33 × 7 = 231) with additional characters there are 276 distinct orthography.

Table 1. Amharic vowels category

	Front	Central	Back
High	ኢ[i]	እ [ɨ]	ኡ [u]
Mid	ኤ[e]	ኧ [ə]	ኦ [o]
Low		ኣ [a]	

To create a complete inventory of Amharic sounds there are a set of thirty-eight phones, seven vowels, and thirty-one consonants [5]. The consonants are classified as stops, fricatives, nasals, liquids, and semi-vowels. Table 2 shows the first three of the Amharic phone inventory.

[1] https://www.ethnologue.com/language/amh (last accessed on 30.11.2018).

Table 2. A few Amharic orthographic inventories

	ə	u	i	a	e	ɨ	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

2.1 Amharic Morphology

Amharic inflectional morphology exhibits addition of prefixes, suffixes, and modifications of root words. A single Amharic word could give hundreds of morphologically inflected different form of words. This morphological richness of the language increases the size of lexicons in speech recognition. The use of morphemes as a subword unit for Amharic speech recognition system is shown on [8, 9, 11] however, there is problem of out of vocabulary (OOV) morphemes. It is practically difficult to use the Amharic rule-based morphological analyzer like HornMorpho[2] for speech recognition purpose. Therefore, researches usually use Morfessor [25] to automatically segment words into morphemes. The tool allows supervised, semi-supervised, and unsupervised training statistical approaches. In this paper, unsupervised training method has been used to prepare morpheme-based corpus. The comparison of morpheme and syllable-based speech recognition model is presented in Sect. 5.

2.2 Amharic Syllabification

Syllable is a unit of sound composed of a central peak of sonority (usually a vowel (V)), and the consonants (C) which cluster around this central peak. Syllabification is the task of segmenting words whether spoken or written into syllables. Technically, the basic elements of syllables are Onset (the first phone in the sequence) and Rhyme (the remaining sequence of phones), which includes nucleus (central peak of sonority) and Coda (the remaining consonants other than the onset) [26]. A syllable can be described by a series of grammars such as consonant-vowel-consonant (CVC) sequence or onset, nucleus & coda (ONC).

Amharic is a syllabic language in which every orthography represents consonant-vowel assimilation. However, not all syllables in Amharic follow the CV sequence represented by the graphemes. Instead, Amharic syllables may follow various patterns, such as V, VC, CV, and CVC, including possible consonant clusters and gemination. Moreover, Amharic orthography did not show epenthetic vowel & geminated consonants that make it challenging to perform syllabification simply following the templates.

A novel syllabification algorithm for Amharic has been shown in [12]. In the paper, acoustic evidence, Amharic syllable template (V, VC, CV, VCC, CVC and CVCC [26]) and the well-known linguistic syllabification implementation principles namely,

[2] https://github.com/hltdi/HornMorpho.

maximum onset and sonority hierarchy principles, have been used to develop a rule-based syllabification algorithm. The algorithm considered gemination and the irregular nature of Amharic epenthesis vowel (ɨ). In this paper, the algorithm has been re-implemented in python with minor improvements. The algorithm is used to prepare syllable-based text corpus for the experiments.

3 Speech and Text Corpus Preparation

Speech recognition research in major languages such as English, German and Chinese has been conducted since 1950s. However, for low-resource languages such as Amharic, there are only a few attempts as it is mentioned in Sect. 1. There are only 20 h of speech data available [10] for the language, which is very less data to develop a better speech recognizer system. It is also challenging to develop Amharic speech-to-speech translation and spoken content retrieval systems [13, 15].

Collecting and preparing a very large speech corpus suited for the development of speech recognizer is costly and labor-intensive task. In this paper, we have prepared approximately 90 h of speech corpus from audiobooks and radio show archives with word and syllable-based transcription. The corpus is merged with the existing dataset and partitioned into training and evaluation set which is made publicly available[3]. An effort has been made to better estimate the number of speakers and age range in the audiobooks and radio show subsets, since we could not found such details.

There are two alternatives in preparing speech corpus. The first alternative is collecting text corpus and ask the native speakers of the language to read the text while recording. The other alternative is finding a variety of prerecorded and transcribed speech and preprocess it for the development of speech recognizer. In this paper, the second alternative is used. However, very few audiobooks and transcribed speech found, which limited the size of the corpus prepared. We have also used publicly available radio program archives. Table 3 provides a summary of all subsets in the corpus.

Table 3. Amharic speech corpus subset summary

Subset	Hours	Gender		Age	#Sentences	#Tokens
		Male	Female			
Existing	20	70	54	18–40	11234	109125
Audiobooks	81	40	–	18–40	22026	339342
Radio	9	30	20	18–50	2780	50208

3.1 Audio Segmentation

For segmenting the audiobooks and the radio show archives, Audacity[4] open source tools have been used. The segmentation process was semi-automatic. Since most of

[3] http://www.findke.ovgu.de/findke/en/Research/Data+Sets/Amharic+Speech+Corpus.html.

[4] https://www.audacityteam.org/.

speech recognition toolkits expect relatively shorter utterances, the average length of the segments is made 14 s. To align the text and spoken sentence command line tools and manual effort has been made. The preprocessing step includes fine-tuning such as removing non-speech contents, removing long silences, and correcting the audio samples using audio processing tools. The sampling frequency for each subset is normalized to 16 kHz with sample size of 16 bits, 256 kbs bitrate with mono channel.

Finally, the corpus is merged with the existing 20 h of data which contains varieties of speakers based on gender, age and dialects. The summary of the dataset could be found in [10].

3.2 Text Preprocessing

After aligning the text with the speech, numbers are converted into equivalent Amharic text as it is spoken in the recordings. Punctuation marks, foreign words, special characters, and symbols have been eliminated, abbreviations are also expanded manually. For some preprocessing tasks, simple python script has been used. For the language model (LM) preparation, CACO the 1.39 million (M) Amharic sentence from [24] has been used. The corpus is merged with our domain-specific text for speech recognition task, which makes it 1.4 M sentences. The text is converted into morphemes for morpheme-base LM using Morfessor 2.0 and into syllables using syllabification algorithm mentioned in Sect. 2.2 for syllable-based LM.

4 Amharic Speech Recognition System

To solve the general speech recognition problem there are three basic modeling approaches, namely, Hidden Markov model (HMM) [1, 16], hybrid Deep Neural Network with HMM model (DNN-HMM) and end-to-end [17, 19, 20] or all neural model. HMM-based automatic speech recognition is a very popular and successful one [16], nevertheless more recently deep neural network (DNN) is becoming state-of-the-art [21, 22].

Since Amharic suffer from lack of standard dataset it is not feasible to go for all neural model. However, in this paper, we have demonstrated the development of a syllable-based DNN-HMM model on the subset of the corpus. In this work, we have tried to balance the benefit of DNN, the use of less dataset and the advantage of getting n-best recognition result. Getting more than one best result is beneficial especially for indexing in spoken content retrieval [2, 3, 15, 18].

In HMM-based ASR, the aim of the system is to find the most likely sentence $\mathrm{W} = \{\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3, .., \mathrm{w}_t\}$ (word sequence) as it is shown in the Eq. (1) which transcribes the speech audio $\mathrm{O} = \{\mathrm{o}_1, \mathrm{o}_2, \mathrm{o}_3, \ldots, \mathrm{o}_t\}$ (acoustic observation).

$$\mathrm{W} = \underset{W}{argmax}\ \mathrm{P}(\mathrm{W}|\mathrm{O}) = \underset{W}{argmax}\ \mathrm{P}(\mathrm{O}|\mathrm{W})\mathrm{P}(\mathrm{W}) \tag{1}$$

Given the phone set, lexicon and the audio files the HMM generates the probability of pronunciation and particular observation sequence given a state sequence which is also referred to as Acoustic model. In the training phase, all the models including the

language model are represented as a weighted finite state transducer (WFST) and they become composed to form one large WFST graph.

4.1 Acoustic Modeling

Acoustic model, P(O|W), represents the relationship between an audio signal and the phonemes or other linguistic units that make up speech. The model learned from a set of audio recordings and their corresponding transcripts [1]. A simple 3-state HMM with its transition probabilities a_{ij} and output probabilities $b_i(o_t)$ is illustrated in Fig. 1.

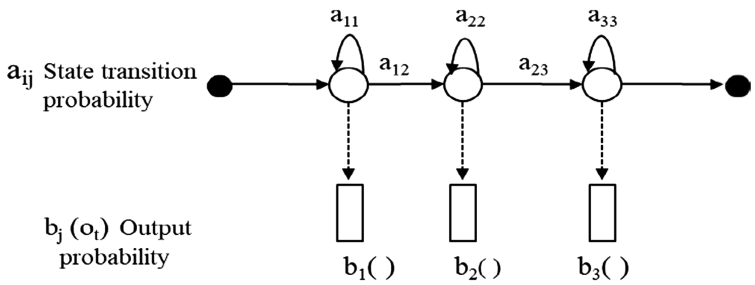

Fig. 1. A simple left-to-right 3-state HMM

Each states capture the beginning, central and ending parts of a phone. In order to capture the articulation effects, triphone models are preferred to context-independent phone models. A mixture of multivariate Gaussian probability distribution functions represented the emission probabilities. The parameters of Gaussian distributions estimated using the Baum-Welch algorithm [16]. In the decoding phase, the dynamic programming Viterbi algorithm is used to get the most probable speech unit sequence (syllable, morpheme or word) sequence from the graph generated.

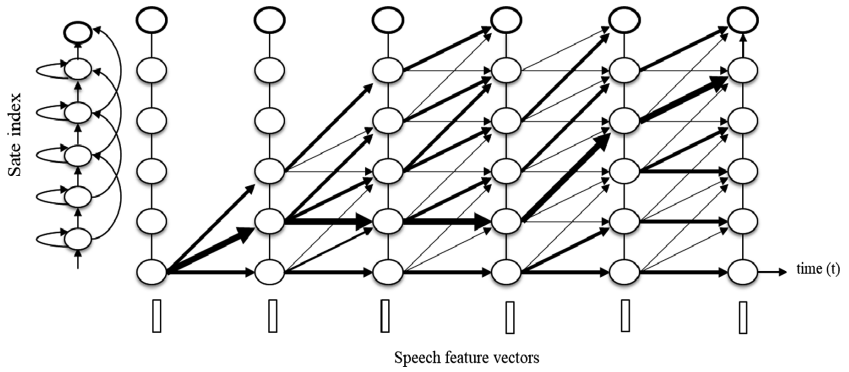

Fig. 2. A Viterbi algorithm for speech unit recognition

As shown in Fig. 2, this algorithm can be seen as finding the best path through a matrix where the vertical dimension represents the states of the HMM and the horizontal dimension represents the frames of speech (i.e. time). Each small circles in the picture represents the probability of observing that frame at that time and each arrow between circles corresponds to a transition probability. Instead of summing over all possible state sequences, we just consider the most likely path which can be achieved by changing the summation to a maximization in the recursion. The score for state j, given the input at time t is computed using Eq. (2).

$$P_j(t) = \max_i \left[P_i(t-1)a_{ij}b_j(t)\right] \tag{2}$$

The paths are grown from left-to-right column-by-column. We need to keep track of the states that make up this path by keeping a sequence of back-pointers. At time t, each partial path is known for all states i, finally we backtrack to find the state sequence of the most probable path. An interesting detail of the application of HMM in speech recognition can be found in [1, 16].

4.2 Language Model

Language model, P(W), it is a probabilistic model used to guide the search algorithm (predict next word given history). It assigns a probability to a sequence of tokens to be finally recognized. The most common modeling approach is the N-gram model $P(w_N|w_1, w_2, \ldots, w_{N-1})$ but recurrent neural network (RNN) is also used as a modeling approach [23]. In this paper, two language models are prepared using a subset of the CACO text corpus mentioned in Sect. 3.2. The first language model is a 1.4 M morpheme-base 5-gram LM and the other is 73k syllable-based 5-gram LM.

4.3 Pronunciation Model

Pronunciation model (lexicon model), P(W|L), forms the bridge between the acoustic and language models [1]. Prior knowledge of language mapping between words and the acoustic units (phoneme is most common). Two different lexicons are prepared for our experiment. The first one is prepared by selecting the most frequent 51k morphemes from the morpheme-based language model text. The second syllable-based lexicon is prepared in the same way by selecting only 16.7k unique syllables from the syllable-based language model corpus.

5 Experiments

The acoustic features extracted for our experiments consist of 13 dimensional Mel Frequency Cepstral Coefficient (MFCC), with their first- and second-order derivatives. A window size of 25 ms with an overlap of 10 ms has been used in the estimation of

the MFCCs. The acoustic models have been trained and tested using Kaldi[5], one of the most widely used open source speech recognition toolkit.

All the language models mentioned in Sect. 4.2 are generated using KenLM[6] statistical language modeling toolkit. The language models are smoothed with modified Kneser-Ney smoothing technique.

5.1 Morpheme-Based System

For the morpheme-based system monophone and triphone models have been experimented in the Kaldi toolkit. The pronunciation dictionary consists of 51k most frequent morphemes described in Sect. 4.3 is used. Moreover, the pronunciation dictionary has been prepared as explained in Sect. 2.1.

In all the models, a 3-state left-to-right HMM topology is used. The monophone model is a context-independent HMM model which does not consider the neighboring phones in the acoustic modeling. The alignment from the monophone model is used as input to the triphone HMMs. Unsurprisingly, the monophone model has worse performance than both the triphone and hybrid DNN-HMM model. Table 4. shows summary of morpheme error rate for each models.

Table 4. Morpheme-based model system performance

Model	Morpheme error rate (MER) %
GMM-HMM monophone	70.97
GMM-HMM triphone	56.36
DNN-HMM triphone	44.62

5.2 Syllable-Based System

As it has been indicated in Sect. 2.2, Amharic has six syllable templates [26]. However, researchers have been considering only the CV syllable template [7, 8]. Moreover, epenthesis vowel and gemination are not handled in those research works [9]. In this paper, all the six syllable templates, as well as epenthesis vowel has been realized using the Amharic syllabification algorithm.

The experimental setup for syllable-based system is the same as morpheme-based system explained in Sect. 5.1, except the lexicon and language model is prepared from syllable units. The lexicon contains only 16.7k syllables and the language model is prepared using 73k syllable-based sentences, which is a small subset of the text corpus. A hybrid DNN-HMM model is experimented with similar setup used in the morpheme-based model.

In the DNN-HMM model, the GMM-HMM alignment from the triphone model is passed into a simple feedforward network (vanilla network with tanh nonlinearities

[5] https://github.com/kaldi-asr/kaldi.git.

[6] https://kheafield.com/code/kenlm/.

adapted from Kaldi script). The network architecture has been built with only 300 hidden layer dimension, 3 hidden layers, minibach size of 128, with initial learning rate 0.04 and final learning rate 0.004. The model is trained for 15 × 2 plus extra 2 × 5 epochs which is 40 epochs in total.

The syllable-based system performed better in all the models even with fewer data in the language model. Moreover, the OOV using syllable units is only three, which is extremely low compared with the morpheme-based system. Table 5 shows a summary of syllable error rate for each model.

Table 5. Syllable-based models system performance

Model	Syllable error rate (SER) %
GMM-HMM monophone	38.00
GMM-HMM triphone	16.82
DNN-HMM triphone	14.36

All the model performance shown in Tables 4 and 5 gained using the subset (20 h) of data to compare the performance of the morpheme-based model and syllable-based model.

6 Discussion and Conclusions

In this paper, new Amharic speech corpus is presented and made available for public access. The dataset is semi-automatically segmented and aligned with word and syllable-based transcript in order to make it suitable for speech recognition and spoken content retrieval tasks. Moreover, syllable-based speech recognition and language models are also introduced. Morpheme and syllable-based models are trained using the existing and the newly prepared corpus. The syllable-based models showed a better result compared with all the morpheme-based models even with language model prepared from a relatively small corpus. The syllable-based system showed a negligible amount of OOV syllables compared with the morpheme-based system. The size of the vocabulary required to prepare the pronunciation dictionary is also noticeably reduced when syllable units are used. The DNN-HMM model showed a better result in all the models even though a simple network with less number of epochs and hidden layers are used. The system provides n-best results in the form of a lattice which makes it a good starting point for tasks like lattice indexing for spoken content retrieval which is planned to be evaluated in future work. As a future work we have also planned to go for all neural model using all the subsets of the dataset.

Acknowledgments. The authors would like to thank the DAAD and MoSHE for funding this research work and DW for allowing us to use Amharic radio program audio from their online archive.

References

1. Gales, M., Steve, Y.: The application of hidden Markov models in speech recognition. Found. Trends® Signal Process. **1**(3), 195–304 (2008)
2. Chelba, C., Timothy, H., Murat, S.: Retrieval and browsing of spoken content. IEEE Signal Process. Mag. **25**(3), 39–49 (2008)
3. Larson, M., Stefan, E.: Using syllable-based indexing features and language models to improve German spoken document retrieval. In: Eighth European Conference on Speech Communication and Technology (2003)
4. Getahun, A.: ዘመናዊ የአማርኛ ሰዋስው በቀላል አቀራረብ (Modern Amharic Grammar in a simple approach), Addis Ababa (2008)
5. Baye, Y.: አጭርና ቀላል የአማርኛ ሰዋስው (Short and simple Amharic Grammar). Addis Ababa (2008)
6. Solomon, T.: Automatic speech recognition for Amharic. Ph.D. thesis (2006). http://www.sub.unihamburg.de/opus/volltexte/2006/2981/pdf/thesis.pdf
7. Solomon, T., Wolfgang, M.: Syllable-based speech recognition for Amharic. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Association for Computational Linguistics (2007)
8. Martha, Y., Solomon, T., Wolfgang, M.: Morpheme-based automatic speech recognition for a morphologically rich language-Amharic. In: Spoken Languages Technologies for Under-Resourced Languages (2010)
9. Martha, Y., Solomon, T., Laurent, B.: Using different acoustic, lexical and language modeling units for ASR of an under-resourced language–Amharic. Speech Commun. **56**, 181–194 (2014)
10. Solomon, T., Wolfgang, M., Bairu, T.: An Amharic speech corpus for large vocabulary continuous speech recognition. In: 9th European Conference on Speech Communication and Technology (2005)
11. Michael, M., Laurent, B., Million, M.: Amharic speech recognition for speech translation. Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN (2016)
12. Nirayo, H., Sebsibe, H.: Modeling improved syllabification algorithm for Amharic. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems. ACM (2012)
13. Chelba, C., Timothy, H., Ramabhadran, B., Saraçlar, M.: Speech retrieval. Spoken language understanding: systems for extracting semantic information from speech (2011)
14. Lee, L., et al.: Spoken content retrieval: beyond cascading speech recognition with text retrieval. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **23**(9), 1389–1420 (2015)
15. Larson, M., Gareth, J.: Spoken content retrieval: a survey of techniques and technologies. Found. Trends® Inf. Retr. **5**(4–5), 235–422 (2012)
16. Huang, X., et al.: Spoken Language Processing: A Guide to Theory, Algorithm, And System Development, vol. 95. Prentice Hall PTR, Upper Saddle River (2001)
17. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
18. Can, D., Murat, S.: Lattice indexing for spoken term detection. IEEE Trans. Audio Speech Lang. Process. **19**(8), 2338–2347 (2011)
19. Amodei, D., et al.: Deep speech 2: end-to-end speech recognition in English and Mandarin. In: International Conference on Machine Learning (2016)
20. Bahdanau, D., Chorowski, J., Serdyuk, D., Bengio, Y., et al.: End-to-end attention-based large vocabulary speech recognition. In: ICASSP, pp. 4945–4949. IEEE (2016)

21. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: ICASSP, pp. 4960–4964. IEEE (2016)
22. Kim, S., Seltzer, M. L.: Towards language-universal end-to-end speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4914–4918. IEEE (2018)
23. Mikolov, T., et al.: Recurrent neural network based language model. In: 11th Annual Conference of the International Speech Communication Association (2010)
24. Andargachew, M.G., Binyam, E.S., Michael, G., Andreas, N.: Contemporary Amharic corpus: automatically morpho-syntactically tagged Amharic corpus. In: Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pp. 65–70 (2018)
25. Sami, V., Peter, S., Stig-Arne, G., Mikko, K.: Morfessor 2.0: python implementation and extensions for Morfessor Baseline. Aalto University publication series SCIENCE + TECHNOLOGY, 25/2013. Aalto University, Helsinki (2013)
26. Mulugeta, S.: The syllable structure and syllabification in Amharic. Masters of philosophy in general linguistic thesis. Department of Linguistics, Trondheim, Norway (2001)