Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4167–4171
Marseille, 11–16 May 2020

Large Vocabulary Read Speech Corpora for Four Ethiopian Languages: Amharic, Tigrigna, Oromo and Wolaytta

Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Atinafu, Binyam Ephrem
Addis Ababa University
solomon.teferra, martha.yifiru, michael.melese, hafte.abera, tewodros.abebe,
wondwossen.mulugeta, yaregal.assabie, million.meshesha, solomon.atnafu, binyam.ephrem@aau.edu.et

Abstract

Automatic Speech Recognition (ASR) is one of the most important technologies to support spoken communication in modern life. However, its development benefits from large speech corpus. The development of such a corpus is expensive and most of the human languages, including the Ethiopian languages, do not have such resources. To address this problem, we have developed four large (about 22 hours) speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. To assess usability of the corpora for (the purpose of) speech processing, we have developed ASR systems for each language. In this paper, we present the corpora and the baseline ASR systems we have developed. We have achieved word error rates (WERs) of 37.65%, 31.03%, 38.02%, 33.89% for Amharic, Tigrigna, Oromo and Wolaytta, respectively. This results show that the corpora are suitable for further investigation towards the development of ASR systems. Thus, the research community can use the corpora to further improve speech processing systems. From our results, it is clear that the collection of text corpora to train strong language models for all of the languages is still required, especially for Oromo and Wolaytta.

Keywords: Speech Corpus, Amharic, Tigrigna, Oromo, Wolaytta

1. Introduction

Natural Language Processing (NLP) applications for many human languages is becoming a required factor for the fulfillment of equity in the information access because one should not be discriminated from the use of the communication technologies just because of not speaking the technologically favored languages. NLP applications are useful in facilitating human-machine and even human-human communications that are inter-mediated by machines. One of the NLP applications which facilitates human-machine and human-human communication is Automatic Speech Recognition (ASR).

ASR is the conversion of speech to the corresponding textual representation. It enables a digital machine to extract the information conveyed in speech as a sequence of words. The resulting sequence of words can be sent to any search engine or can be used as the transcription of any human speech. This helps illiterate people who cannot read and/or write, hands busy people, transcribers of legal, medical, meeting (parliamentary for example) speeches.

The need for the ASR technologies and the advancements in powerfull algorithms, such as deep learning, processing power with graphics processing unit (GPU) and availability of large amounts of data have encouraged research and development in ASR. As a result, a considerable amount of research on the development of ASR Systems (ASRS) has been conducted and lots of ASRSs have been developed. So far, however, ASRSs have only been developed for a few of more than 7000 languages in the world.

The development of ASR requires three language resources in a particular language. These are a speech corpus that consists of a large number of audio files along with the text transcription; a pronunciation dictionary of at least all the words in the transcription and a large collection of text for language modeling. Among these language resources, developing a speech corpus for a number of languages is not an easy and cheap task.

Lack of electronic resources for speech and language processing is one of the major aspects considered to categorize languages as under-resourced (Besacier et al., 2014). On the basis of this fact, the Ethiopian languages are under-resourced. On the other hand, digital technologies are now coming up with speech input interfaces that demand the development of ASRSs, which in turn requires speech corpus. To address this problem, we have developed a speech corpus that consists of about 100 hours of training speech for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta.

The corpora consist of 26.5 hours of Amharic, 22.1 hours of Tigrigna, 22.8 hours of Oromo and 29.7 hours of Wolaytta training speech. Using these corpora we have developed HMM-GMM based ASR that could be considered as a baseline for further use of these corpora for research on speech processing and for further development of ASR systems for any of these languages. The procedure we applied for the development of these corpora may also serve as a model for collecting additional corpora in many of the more than 70 remaining Ethiopian languages.

In section 2. we present a brief description of the languages considered in this paper. The detailed analysis of the speech corpora and how they are developed are presented in section 3. The development of the baseline ASR systems and the results achieved are presented in section 4. Conclusions and future directions are presented in section 5.

2. Ethiopian languages

The languages considered in this work belong to Semitic (Amharic and Tigrigna), Cushitic (Oromo) and Omotic (Wolaytta) language families. All of these languages have a considerable number of native speakers. Amharic is spoken

by more than 27 million people while Tigrigna is spoken by 9 million people. Oromo and Wolaytta are spoken by more than 34 million and 2 million speakers, respectively (Simons and Fennig, 2017). These languages have different functions in Ethiopia. Amharic, for instance, is the working language of the Federal Government. It also serves as regional working language of some other regional states. Tigrigna and Oromo are working languages in Tigray and Oromiya regional states, respectively. Some of the governmental websites are available in Amharic, Tigrigna and Oromo. Apart from this, they serve as medium of instructions in primary and secondary schools. These languages are available in electronic media like news, blogs and social media. Currently, Google offers a searching capability using Amharic, Tigrigna and Oromo. Furthermore, Google also included Amharic in its translation services recently.

2.1. Phonology

Even if these four languages belong to three different language families, they share a lot of phonetic properties. They share about 70% of their phone sets, including the ejectives tʼ kʼ pʼ tsʼ tʃʼ that are not used in most of the languages in the other parts of the world. Long consonants, geminated consonants, are clearly pronounced and bring semantic difference in these languages. Since analysis of their phonetic relations is not the scope of this paper, we describe only the most prevalent relations between two language pairs.

The Amharic phone set is a subset of the Tigrigna phone set. Amharic has a total of 35 phones (28 consonants and 7 vowels) while Tigrigna has 39 phones (32 consonants and 7 vowels). Tigrigna has four sounds that are not found in Amharic: ʕ, ħ, x and x́. Both languages have seven vowels: ə, u, i, a, e, ɨ, o.

Although they belong to different language families, Oromo and Wolaytta have more phonetic commonalities with each other than commonalities they have with the above mentioned Semitic languages. The major one is their use of long and short variants of the same five vowels. So each of these languages has ten vowels. The other common phonetic feature of these languages is the use of tones which makes both of them tonal languages. Having their own inventory of consonants, Oromo (28) and Wolaytta (27) they share a number of consonants. Of course, there are consonants which are not shared between these languages. The consonants ɲ and x are used in Oromo but not in Wolaytta while the consonant ʒ is not used in Oromo but in Wolaytta.

2.2. Morphology

All the four languages can be considered as morphologically complex. Morphologically, we can categorize them into two. Reflecting their Semitic language morphology, Amharic (Leslau, 2000) and Tigrigna (Tesfay, 2002), make use of the root and pattern system. In these languages, a root (which is called a radical) is a set of consonants which bears the basic meaning of the lexical item whereas a pattern is composed of a set of vowel patterns inserted between the consonants of the root. These vowel patterns together with affixes, results in derived words. Such a derivational process makes these language to be morphologically complex. In addition to the morphological information, some syntactic information are also expressed at word level.

Unlike the Semitic languages which allow prefixing, Oromo and Wolaytta are suffixing languages. In addition, words can be generated from stems recursively by adding suffixes only. The morphology of Oromo and Wolaytta is, however, relatively simpler than Amharic and Tigrigna. This has been reflected in the smaller out of vocabulary (OOV) rate and the growth of their vocabulary size as presented in (Abate et al., 2018).

In all the four languages, nominals are inflected for number, gender, definiteness and case whereas verbs are inflected for person, number, gender, tense, aspect, and mood (Griefenow-Mewis, 2001).

2.3. Writing System

The writing systems of these languages are Ethiopic and Latin. Amharic and Tigrigna are written in Ethiopic while Oromo and Wolaytta are written in Latin. The Ethiopic is a syllabic script where each character represents a consonant and a vowel. This writing system does not distinguish the short and long form of a consonant and the presence and absence of the epentetic vowel and the glottal stop consonant is not marked in the writing.

Oromo and Wolaytta writing system uses Latin. In both languages, the current writers differentiate the geminated and the non-geminated consonants. Similarly, long and short vowels are indicated in their writing system. So the text has a clear and consistent grapheme-to-phoneme (G2P) relations in both these languages.

3. Development of Speech Corpus

To develop any ASR systems one requires all the three language resources, indicated in section 1., in a language. However, only two (Amharic and Tigrigna) of the more than 80 Ethiopian languages have spoken language resource.

3.1. Development of Speech Corpus for Ethiopian Languages

Although different attempts towards the development of ASR for the Ethiopian languages have been made by student researchers as part of their academic requirements, most of their works depend on small sets of speech data. The only works known for the development of standard speech corpora for Ethiopian languages are the development of the medium-sized read speech corpus (Abate et al., 2005) and (Pellegrini and Lamel, 2009) for Amharic and the development of a similar speech corpus for Tigrigna (Abera and Hailemariam, 2018).

To the best of our knowledge, there are no standard speech corpora for all of the Ethiopian languages except the above mentioned two languages. In this work, we have, therefore, developed speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta.

3.1.1. Text Collection and Pre-processing

Following the steps of read speech corpus development, we have first collected and pre-processed a large text corpus from the web and different previously developed text sources for each of these languages. The texts in

Amharic and Tigrigna are mainly from broadcast news domain while the texts in Oromo and Wolaytta are from different domains, including spiritual (Bible) domain. We have then, applied different automatic methods for text pre-processing. As part of the pre-processing, unnecessary links, numbers, symbols and foreign texts have been removed and the following tasks have been performed: character normalization and sentence tokenization.

We have then selected the text transcription using two methods. For Amharic and Tigrigna we have used the algorithm developed by (Abera et al., 2016) that uses characters as units to analyse the phonetic balance and richness of the database. The algorithm is based on the syllabic writing system of Tigrigna and could be used for Amharic that uses the same writing system. Since Oromo and Wolaytta use Latin script, we could not apply the algorithm that is used for Amharic and Tigrigna to analyse phonetic balance and richness of training transcriptions. We have, therefore, considered sentences that are shorter than 20 words to minimize reading difficulties.

Aiming at having at least 100 clean audio files from each speaker, we have assigned 130 utterances for each of the 100 Amharic, Tigrigna and Oromo speakers. Considering the difficulty to get 100 speakers and the probability that a reader may not read all the assigned utterances, we have increased the number of sentences to be read by a speaker for Wolaytta to 150. Distinct set of prompts for each speaker in each language have been randomly selected from the prepared text databases.

3.1.2. Speech Recording

For recording the speech, we have used 6 mobile phones of TECNO K7 on which an Android based speech recording software (Gauthier et al., 2016) has been installed. The software is set up to capture wave from only one channel with a sampling rate of 16kHz and the encoding of 16bit pulse-code modulation (PCM). Using this setup, we have applied a semi-supervised recording approach. The speakers were given a short description of the recording system and allowed to choose a place and time of their convenience to conduct the recording. The recording software displays the text for them one sentence at a time and as they finish reading the recorded audio file is saved. The recording has been done at Addis Ababa University for Amharic, Tigrigna and Oromo while the Wolaytta speech is recorded at Wolaytta Sodo University.

The readers were selected from university students at the department of Tigrigna and Oromo of the Addis Ababa University while the readers of the Amharic speech are students of mainly the school of Information Science of the same university. Consequently, university students are dominant who are younger than 40 years old. The Wolaytta readers are selected from Wolaytta Sodo University based on availability of native speakers of the language, irrespective of their age and educational level (if they can read). In selecting the readers native speakers are considered.

The recording process took place for about 4 months for Oromo and Tigrigna and more than 6 months for Amharic and Wolaytta to complete the whole recording. After the recording process has come to an end we have processed the text and speech data in a way it can be used in the Kaldi ASR development format. As part of this post-processing a lot of audio files have been filtered out. Some of them were empty files, others consist of only a few part of the sentence, still others consist of repetitions of words or phrases and did not pass the alignment step of acoustic model (AM) training. In this paper we are presenting the remaining speech corpora which have been checked for their usefulness by developing a baseline ASR system for each of them.

3.1.3. Details of the Speech Corpora

A close investigation of the corpora that are considered to be usable shows that 98 Amharic, Tigrigna and Oromo speakers and 85 Wolaytta speakers have been recorded. But not all of these speakers have read all the 130 (for Amharic, Tigrigna and Oromo) and 150 (for Wolaytta) utterances. So we show the distribution of the number of audio data per speaker for each language in Table 1.

No. of Utt	Amharic	Tigrigna	Oromo	Wolaytta
80-99	1	1		4
100-120	5	3	7	1
121-125	29	31	36	1
126-130	63	63	55	
131-140				8
141-145				27
146-150				44
Total	98	98	98	85

Table 1: Number of Utterances per speaker

The age and gender distributions of all the speakers is shown in Tables 2 and 3.

Age range	Amharic	Tigrigna	Oromo	Wolaytta
15-19	5			9
20-29	91	30	19	39
30-39	2	67	75	32
Above 40		1		4
Unkown			4	1
Total	98	98	98	85

Table 2: Age Distribution of the speakers

Rs Gender	Amharic	Tigrigna	Oromo	Wolaytta
Male	45	48	49	51
Female	53	50	49	34
Total	98	98	98	85

Table 3: Gender Distribution of the speakers

4. Development of Baseline ASRSs for Ethiopian Languages

The usefulness of the described corpora for building speech processing components was investigated by developing ASR systems for each of the four languages. A brief description of the procedure we followed is presented in subsection 4.1. We consider our results as baselines for further use of the corpora in the development of ASR systems for these languages.

4.1. Lexical, Language and Acoustic Models

The list of words for the development of the training vocabularies have been extracted from the transcription of the training speech by word tokenization. There are 12,328,

11,305, 11,297, and 10,939 sentences in the training transcription of Amharic, Tigrigna, Oromo and Wolaytta, respectively. The size of the vocabularies for each language is given in Table 4. For all the languages we have used the respective training vocabulary both for training and decoding purpose.

Using the syllabic nature of Amharic and Tigrigna writing system, we have generated the pronunciation dictionaries automatically. As indicated in section 2.3., Oromo and Wolaytta exclusively differentiate consonant and vowel variants in their writing system which is used for our Automatic G2P conversion.

For the development of the language models (LMs), we have used the text developed by (Tachbelie and Abate, 2015) for Amharic, and raw texts collected and preprocessed for Tigrigna and Oromo. For Wolaytta, we could not collect more text than the training speech transcription and therefore, the training transcription is used for language modeling.

For each of the languages, we have developed trigram LMs using the SRILM toolkit (Stolcke, 2002). The open vocabulary LMs are smoothed with unmodified Kneser-Ney smoothing techniques (Chen and Goodman, 1996). The LM probabilities are computed only for the list of words in the transcription of the training speech (training vocabulary) using the LM training text that are different in size. The size of the training text, the OOV rate and the perplexity of the LMs with respect to the test sets are given in Table 4.

Languages	LM text in token	Vocab in thousands	OOV	PPL
Amharic	4M	43	26.83	39.63
Tigrigna	4M	32	15.47	107.30
Oromo	1.2M	21	11.73	266.17
Wolaytta	226k	25	9.34	254.90

Table 4: *Language Model Statistics*

For ASR evaluation purpose, we have held out speech data of four speakers for development and four speakers for evaluation sets for each language, except Amharic. Each of the test sets consist of two male and two female speakers that are randomly selected from the total speakers. The remaining speech of 90 speakers for Tigrigna and Oromo and 78 speakers for Wolaytta is used for training the AMs. For Amharic, we have used the test sets of the Amharic read speech corpus (Abate et al., 2005) that consists of 10 speakers for development and evaluation test sets each. Thus, all the Amharic speech read by 98 speakers has been used for the training of the AMs.

All the AMs were built in the same way using Kaldi ASR toolkit (Povey et al., 2011), one of the most widely used open source speech recognition toolkits. We have built context dependent HMM-GMM based AM for each language using 39 dimensional mel-frequency cepstral coeffcinents (MFCCs). The AM uses a fully-continuous 3-state left-to-right HMM. Then we did Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation for each of the models. Then Speaker Adaptive Training (SAT) has been done using an offline transform, feature space Maximum Likelihood Linear Regression (fMLLR). The performance of the systems with respect to the evaluation test sets is given in Table 5.

Languages	Amharic	Tigrigna	Oromo	Wolaytta
WER in %	37.65	31.03	38.02	33.89

Table 5: Performance of the Baseline ASRs

As we can observe from Tables 4 and 5, the morphological nature of the languages have a significant impact on the performance of the respective ASR systems. Although the Amharic LM benefited from the clean and in domain text used for training as reflected in the lowest perplexity it achieved, the higher OOV rate of the decoding vocabulary resulted in a high WER. The Tigrigna LM has higher perplexity than that of the Amharic LM resulting from the LM training text collected from different domains including Bible text. However, the decoding vocabulary has lower OOV rate than that of Amharic. That resulted in a better performance of the ASR system. Although we used a poor quality LM training text for Oromo and Wolaytta (domain mixed for Oromo and very small for Wolaytta) which is also reflected in the high perplexities of the LMs, the ASR systems benefited from the lower OOV rate which in turn reflects the lower morphological complexity of these languages.

5. Conclusions and Future Works

In this paper, we presented the development of four speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. Each of the corpora consists of speech of more than 22 hours. We also presented the performance of the baseline ASR systems we have developed using the respective corpora. The WERs we achieved showed that the corpora are suitable for further investigation and development of ASR systems for these languages. The data will be shared with the research community to foster further development of speech processing components. We recommend the collection of more text corpora to train better LMs for all of the languages especially for Oromo and Wolaytta that do not have large text collected for language modeling purpose.

Since the corpora are developed by a project financed by the Addis Ababa University, we intend to make them freely available for research purpose via language resource distributors like European Language Resources Association (ELRA) as per the rules and regulation of the Addis Ababa University.

6. Acknowledgment

We would like to express our gratitude to the Addis Ababa University for funding the project and all the readers whose speech are included in the corpora. We also appreciate Hafte Abera and Seifedin Shifaw for collecting raw text for Tigrigna and Oromo, respectively. We are also thankful to the Cognitive Systems Lab (CSL) of the University of Bremen where we have completed all the post processing works and the development of the baseline ASR systems. Above all, we are grateful to the constructive comments and a rich experience we have got from Prof. Tanja Schultz.

7. Bibliographical References

Abate, S. T., Menzel, W., and Tafila, B. (2005). An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 1601–1604.

Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., Assabie, Y., Abera, H., Ephrem, B., Abebe, T., Tsegaye, W., Lemma, A., Andargie, T., and Shifaw, S. (2018). Parallel corpora for bilingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Abera, H. and Hailemariam, S. (2018). Design of a Tigrinya language speech corpus for speech recognition. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 78–82, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Abera, H., Nadeu, C., and Mariam, S. H. (2016). Extraction of syllabically rich and balanced sentences for tigrigna language. In *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, Jaipur, India, September 21-24, 2016*, pages 2094–2097.

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

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

Gauthier, E., Blachon, D., Besacier, L., Kouarata, G., Adda-Decker, M., Rialland, A., Adda, G., and Bachman, G. (2016). Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 381–382.

Griefenow-Mewis, C. (2001). *A Grammatical Sketch of Written Oromo*.

Leslau, W. (2000). *Introductory grammar of Amharic*. Porta linguarum orientalium, neue Serie, Bd. 21. Harrassowitz, Wiesbaden.

Pellegrini, T. and Lamel, L. (2009). Automatic word decompounding for ASR in a morphologically rich language: Application to amharic. *IEEE Trans. Audio, Speech & Language Processing*, 17(5):863–873.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002*, pages 901–904.

Tachbelie, M. Y. and Abate, S. T. (2015). Effect of language resources on automatic speech recognition for amharic. In *AFRICON 2015, Addis Ababa, Ethiopia, September 14-17, 2015*, pages 1–5.

Tesfay, T. Y. (2002). *A modern grammar of Tigrinya*. Tipografia U. Detti, Rome.