Design of a Tigrinya Language Speech Corpus for Speech Recognition

Hafte Abera
Addis Ababa University
Addis Ababa, Ethiopia
hafte.abera@ aau.edu.et

Sebsibe H/Mariam
Addis Ababa University
Addis Ababa, Ethiopia
sebsibe2004@gmail.com

Abstract

In this paper, we describe the first Tigrinya Language speech corpus designed and developed for speech recognition purposes. Tigrinya, often written as Tigrigna (ትግርኛ) /tɪˈgrinjə/ belongs to the Semitic branch of the Afro-Asiatic languages and shows characteristic features of a Semitic language. It is spoken by ethnic Tigray-Tigrigna people in the Horn of Africa. This paper outlines different corpus designing processes and related work on the creation of speech corpora for different languages. The authors also provide procedures that were used for the creation of a speech recognition corpus for Tigrinya, an under-resourced language. One hundred and thirty native Tigrinya speakers were recorded for the training and test datasets. Each speaker read 100 texts, which consisted of syllabically rich and balanced sentences. Ten thousand sets of sentences were used, which contained all of the contextual syllables and phones of Tigrinya.

1 Introduction

Speech corpus is defined as a collection of speech signals that is accessible in computer readable form, and has an annotation, metadata and documents to allow re-use of the data in house or by scientists (Gibbonet et al., 1997). Speech corpus is one of the fundamental requirements for developing speech recognition and synthesis systems and to analyze the characteristics of speech signals. Moreover, for phonetic research, speech corpus can also provide diverse and accurate data to help researchers find the rules of languages. The main reason for preparing a Tigrinya speech corpus is, as a first step, to explore the possibility of developing a Tigrinya speech recognition system (Li & Zu, 2006). Tigrinya, often written as Tigrigna (ትግርኛ) /tɪˈgriːnjə/belongs to the Semitic branch of the Afro-Asiatic languages and shows characteristic features of a Semitic language (Tewolde, 2002; Berhane, 1991). It is spoken by ethnic Tigray-Tigrigna people in the Horn of Africa.

Speech corpora design is one of the key issues in building high quality text to speech recognition systems (mostly read speech). Radová (1998) pointed out that most speech corpora contain read speech, either for practical reasons because annotating non-read speech is more difficult, or simply because the intended application or investigation requires read speech. Due to the first reason a read speech corpus is prepared and used for this work.

The preparation of any type of speech corpus is normally a project on its own and handled on the basis of an agreement between corpus producers and corpus users. However, in cases like this study, where the required corpus is not available, speech recognition experiments are conducted on the newly produced corpus. The advantage in the latter case is that the corpus is produced with full and specific knowledge of its intended use.

According to Li & Yin (2006), the development of speech corpus has created new problems: (1) many corpora have been established, and much money and time have been put into their technology; and (2) these corpora are difficult to share among different affiliations. The foremost cause of this problem is the lack of general specifications for corpus collection, annotation, and distribution. In order to solve this problem, standardization research on speech corpora is necessary and specifications should be stipulated (Li &Yin, 2012).

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 78–82
Santa Fe, New Mexico, USA, August 20, 2018.

2 Benchmarks and procedure for speech corpus construction

The Tigrinya Speech Recognition Corpus has been designed to satisfy a set of guidelines, which specify the required quality of speech data and the proportional distribution of data with different speaker characteristics. In this paper the authors provide a brief overview of the criteria defined for the speech corpus from Li & Zu (2006). The criteria are explained in the following sub-sections.

2.1 Specification of speakers

Any speech corpus should be representative of speakers of both genders in equal proportions and contain speech from speakers of different ages. In this work, the speakers were grouped into six age groups, as seen in Table 1: from 18 to 22 (up to 19%), from 23 to 27 (at least 25%), from 28 to 32 (at least 25%), from 33 to 37 (at least 15%), from 38 to 42 (at least 13%), and from 43-70 (at least 4%). The upper boundary is relatively small, since we did not have enough speakers to cover that age group. The researchers excluded the speech of children and people older than 70, because the speech characteristics in these groups are quite different and require training of separate acoustic models.

Age Range	**Training set**		**Test Set**		**Adaptive data set**		**Total**
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	
18-22	10	9	1	1	3	1	25
23-27	12	10	1	2	4	3	32
28-32	14	12	1	2	1	2	32
33-37	8	8	1	1	1	1	20
38-42	9	6	0	0	1	1	17
43- 70	1	1	0	0	1	1	4
Total	54	46	4	6	11	9	130

Table 1: Age and Sex Distribution of the Readers

2.2 Specification of corpus design

After having the text corpus, the next important step in preparing a read speech corpus is recording the speech. In the recording of selected sentences, each speaker is asked to read exactly what is presented to him or her. The text to be read is presented on a mobile phone. In the read speech recording, the degree of control is very high. For example during the recording, each utterance of the speaker can be checked directly for errors, and if an error is found, the speaker is asked to re-read the text.

The recordings of the Tigrigna speech corpus were done in an office environment, as well as outside the office. The text was read by 130 native Tigrigna speakers. For recording purposes, the mobile application Lig-Akuma, which displays one sentence at a time for the speaker to read, was used. The entire recording was done in the presence of a researcher. The speaker was first explained the purpose of the project and instructed what to do. The recording session was controlled by the researcher, which included running the recording program, starting and stopping the recording session, playing back the recorded speech, re-recording the sentences (if required), and moving to the next sentence. Every speaker was instructed to start the recording when he or she was ready. After each reader's session was finished, all the utterances were listened to both by the reader and the researcher for possible corrections.

The Tigrigna speech corpus was designed according to best practice guidelines established for other languages. Standard speech corpora, such as the Amharic speech corpus (Abate et al., 2005), consist of a training set, a speaker adaptation set, and test sets. To make it comparable with commonly used standard corpora, the Tigrinya corpus was designed to contain the same components.

The recording training set consists of a total of ten thousand different sentences. The training set was read by 100 speakers of the different Tigrai dialects. As already mentioned, Table 1 shows the age and sex distribution of all speakers. Due to time constraints, it was difficult to keep the age and sex balance in the speaker sample.

Each test and speaker adaptation set was read by 10 and 20 speakers, respectively, from different dialects. For the test sets, different sentences were selected for each speaker. Table 2 shows the number of sentences that have been automatically selected from the text database and read for the collection of speech data for the corpus.

Data-Set	Number of selected sentences	Number of recorded sentences	Duration (hours)
Training set	10,000	10,000	17:57
Speaker Adaptation set	69	69	2:30
Development test set	1,000	600	2:10
Evaluation test set	1,000	400	1:20

Table 2: Elements of Tigrinya Speech Corpus

2.3 Specification of recording

Data collection was carried out using an improved version of the Android application Lig-Aikuma, developed by Steven Bird and colleagues (Blachon et al., 2016). The resulting app, called Lig-Aikuma, runs on various mobile phones and tablets and proposes a range of different speech collection modes (recording, re-speaking, translation and elicitation). Lig-Aikuma's improved features include a smart generation and handling of speaker metadata as well as re-speaking and parallel audio data mapping.

For every mode, a metadata file was saved with the recording file. Metadata forms are filled in before any recording. In addition, metadata have been enriched with new details about the languages (language of the recording, mother tongue of the speaker, other languages spoken) and about the speaker (name, age, gender, region of origin). Moreover, in order to save time, a feature saves the latest metadata as a session and uses it to preload the form the next time it is necessary. The files are now named using the specific following format: DATE-TIMEDEVICE-NAME LANG. As an example 2016-02-07-01-32-18_HUAWEI-HUAWEI Y625-U32_tir is the name of a recording made on February, 07, 2016, at 1pm (01:32:18), in Tigrinya, on a HUAWEI device.

Similarly to the general trend of speech recognition system development (e.g., Abate et al., 2005; Blachon et al., 2016; De Vries et al., 2011), all speech data were sampled in 16 KHz and 16 bits allocated per sample. In order to be able to develop speech recognition systems that can be executed in different environments and that can be used for wide purposes, we should be able to build speech acoustic models that are robust, as well as noise representative to different environments (Li &Zu, 2006). Therefore, the Tigrinya speech recognition corpus has been developed to include speech data with different types of background noise (office, street, in-car, etc. noise) and with different signal-to-noise ratios (SNR). The majority of the data, however, contain a relatively low level of noise (the SNR being between 20-85dB). Overlapping speech segments (with two or more simultaneous speakers) are not included in the corpus. The Tigrigna Speech Corpus consists of utterances from Tigrinya literary (formal) language, however, pronounced by a variety of speakers, including speakers with different dialect from both the Southern and Northern Tigrinya dialects (Berhane, 1991).

2.4 Specification of annotation

Following recent research on other language speech recognition corpora (Abushariah et al., 2012; Radová, 1999; Arora et al., 2004) the corpus has been designed to be phonetically balanced in order to be representative of natural speech and phonetically rich so that the trained acoustic models would efficiently generalize over different speakers with different characteristics. Abera et al. (2016) previously showed how to analyze and create corpora with respect to syllabic richness and balance. An example of an orthographically annotated utterance is given in Figure 1.

2.5 Availability

Thus far the corpus has been used for our research in developing an automatic speech recognizer for Tigrinya. As our research is at its last phase, we have the intention of making the corpus available for researchers and developers by means of a third party.

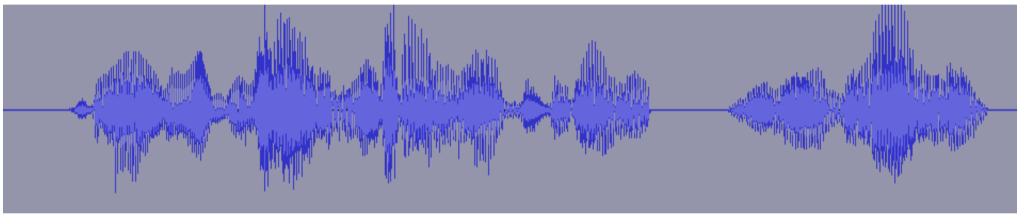

Figure 1: Example of the Phonologically Annotated Utterance ፃዕሪ ግብፅን ንግድ ኢትዮጵያን እቶም ነዊሕ ዝቑኣመቶም ሰብ ሳባን ናባኻ ኺሐልፉ ናትካውን ኪኾኑ እዮም "Tsaeri gbtsn ngd etyopyan etom newih zquametom seb saban nabaKa KiHalf Natkawn kiKuinu eyom"

3 Conclusion and future work

In this paper the authors presented the overall design of the Tigrinya Speech Recognition Corpus. The corpus is the first database available to train and test an Automatic Speech Recognition system, which takes into account dialectal variation in Tigrinya. The corpus consists of 17.57 hours of speech audio data set for training and 3.3 hours of phonetically annotated speech audio data for development and evaluation. The corpus is both phonetically rich and balanced. This paper also described the reasoning behind different choices made during the development of the corpus. We believe that this paper can serve as a guideline for other researchers, who are developing, or want to develop speech recognition corpora for under-resourced languages.

The Tigrinya Speech Recognition Corpus in the years to come will be an asset for further research in speech recognition in general, as well as speech synthesis of Tigrinya – a language that did not have a dedicated speech recognition corpus before. In addition, the speech corpus should be able to play a crucial role in linguistic research, such as comparing the differences in pronunciation made by men and women. Additional experiments with this corpus are anticipated to improve the performance of our speech recognition system.

References

Solomon Teferra Abate. 2006. *Automatic speech recognition for Amharic*. Ph.D. Thesis. University of Hamburg. Hamburg.

Solomn Teferra Abate, Wolfgang Menzel, and Bairu Tafila. 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In Eurospeech, 9th *European Conference on Speech Communication and Technology*, pages 1601-1604, Lisbon, Portugal.

Hafte Abera, Climent Nadeu, and Sebsibe H. Mariam. 2016. Extraction of syllabically rich and balanced sentences for Tigrigna language. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, pages 2094-2097, Jaipur, India.

Mohammad AM Abushariah, Raja N. Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman O. Khalifa. 2012. Phonetically rich and balanced text and speech corpora for Arabic language. *Language resources and evaluation*, 46(4), 601-634.

Karunesh Arora, Sunita Arora, Kapil Verma, and Shyam Sunder Agrawal. 2004. Automatic extraction of phonetically rich sentences from large text corpus of Indian languages. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, pages 2885-2888, Jeju, Korea.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app. *Procedia Computer Science*, 81, 61-66.

Nic De Vries, Jaco Badenhorst, Marelie H. Davel, Etienne Barnard, and Alta De Waal. 2011. Woefzela-an open-source platform for ASR data collection in the developing world. In *Proceedings of INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 3177-3180, Florence, Italy.

Dafydd Gibbon, Roger Moore, and Richard Winski. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin New York.

Ai-jun Li and Zhi-gang Yin. 2007. Standardization of speech corpus. *Data Science Journal*, 6,: S806-S812.

Ai-jun Li and Zu, Yiqing. 2006. Corpus design and annotation for speech synthesis and recognition. *Advances in Chinese spoken language processing*: 243-268.

Vlasta Radová. 1998. Design of the Czech Speech Corpus for Speech Recognition Applications with a Large Vocabulary. In *Text, Speech, Dialogue. Proc. of the First Workshop on Text, Speech, Dialogue*. Brno, Czech Republic, pages 299-304.

Vlasta Radová, Petr Vopálka, and P. Ircing. 1999. Methods of Phonetically Balanced Sentences Selection. In *Proceedings of the 3rd Multiconference on Systemics, Cybernetics and Informatics*, pages 334-339, Orlando, USA.