Large Vocabulary Read-Mode Speech Corpora for Low-Resourced Ometo Languages: Gamo, Gofa, Dawuro and Wolaita

N. S. Sundado* NEBIYU.SIMON@WSU.EDU.ET
Department of Electrical and Computer Engineering, Wolaita Sodo University, Ethiopia

M. M. Woldeyohannis* MICHAEL.MELESE@AAU.EDU.ET
School of Information Science, Addis Ababa University, Ethiopia

A. E. Kurka AKLILU.ELIAS@WSU.EDU.ET
Department of Information Technology, Wolaita Sodo University, Ethiopia

Abstract

Speech is a fundamental mode of human communication and has also become a popular way for people to interact with machines through the use of speech technology. Automatic Speech Recognition (ASR) is one of speech technologies which transcribes speech to its corresponding text using numerous techniques and facilitates communication between human and electronic devices. To make it a real large amount of speech dataset parallel with its transcription is necessary. However, developing a large amount of corpora through collection and pre-processing is very expensive for many languages, including Ometo languages. This is mainly because they are classified as low-resource languages, lacking sufficient linguistic data and technological resources. In order to solve the problem of data scarcity for Ometo languages: Gamo, Gofa, Dawuro, and Wolaita, we have developed large speech corpora of 24.3511 hr with its corresponding transcription for four Ometo languages. Then, we have developed ASR systems for each language to verify the usability of the corpora using a deep learning technique. We have achieved WER of 72.00%, 57.94%, 62.22%, 64.71% for Gamo, Gofa, Dawuro and Wolaita languages, respectively. In order to demonstrate that the corpora are appropriate for additional research toward the creation of ASR systems, we present the corpora and the baseline ASR systems we have constructed in this study. The corpora can therefore be used by researchers to improve speech processing systems.

Keywords: Automatic Speech Recognition, Low-Resource Languages, Ometo Languages, Deep Learning

1. Introduction

Modern human societies depend heavily on spoken language. Simply, speech is a means of language based communication between individuals. Now, it has also evolved into a means of communication between people and machines (or computers) in the way that is natural, adaptable, effective, and convenient while allowing hands and eyes to be free exploiting intelligent tool with speech interface for the aim of simplifying people's daily life (Fanta, 2010; Liang and Yan, 2022).

Applications of natural language processing (NLP) for a wide range of human languages are becoming necessary for the realization of fairness in the access to information, as no one should be excluded from using communication technologies solely because they do not speak a language that is preferred by the technology. Applications of natural language processing (NLP) are helpful in enabling machine-mediated human-machine and even human-human conversations (Abate et al., 2020). In order to facilitate the communication between people and computer, automatic speech recognition is being applicable as recent technology (Saksamudre et al., 2015).

It is done through automatic transcription of spoken words into text for other systems, including information retrieval systems, allows for text input. It mainly seeks to translate the audio material into the appropriate text. The text is then processed further through human-computer interaction, including translations into several languages and hand-talk, among other things. This can help those who are illiterate and

* These authors contributed equally

incapable of reading or writing, as well as busy individuals, translators of speeches from meetings (including legislative ones), and professionals in the legal and medical domains (Abate et al., 2020).

Therefore, in order to achieve the necessary goal, there has been a recent increase in emphasis paid to the rapid improvement of highly performing automatic speech recognition (ASR) systems for a range of languages as they move through distinct phases of development. It became essential for all human languages since voice is difficult for machines to process directly in human-machine communication. This has led to the development of multiple Automatic Speech Recognition Systems (ASRSs) in different human languages. It has taken a lot of study and development to accomplish this (Tachbelie et al., 2020; Mohamed et al., 2011). However, only a small portion of the world's more than 7100 languages got attention for the application of automatic speech recognition due to many languages are suffering from the data scarcity.

Despite its improved result, DNN, the latest technique for ASR, needs a lot of data to have good result. While there is plenty of data in case of English language which does not hold true for many other languages like Ethiopian languages including Ometo languages (Gamo, Gofa, Dawaro and Wolaita). This truth makes sure that the effectiveness in the development of ASR systems for human languages is the availability of speech corpora for languages. However, it is challenging to find sufficient corpora for a sizable number of human languages, many of which are known to be under-resourced (Tachbelie et al., 2020; Müller and Waibel, 2015). One of the main factors used to classify languages as under-resourced is the absence of electronic resources for speech and language processing (Besacier et al., 2014).

The majorities of Ethiopian languages, including Ometo languages are under-resourced and are among the language families that are not gaining from the advancement of spoken language technologies. To address the issue of scarcity of voice corpora for under-resourced Ometo languages, we attempted to collect speech corpora for Gamo, Gofa, Dawaro and Wolaita with selected age groups and both genders and also to develop baseline models. We have developed a speech corpus that consists of about 24.3511 hours of speech for four Ometo languages: Gamo, Gofa, Dawaro and Wolaita. The speech corpora consist of 5.535 hours of Gamo, 6.0358 hours of Gofa, 6.8103 hours of Dawuro and 5.97 hours of Wolaita. With utilizing these corpora, we are able to develop deep learning-based ASR and obtain WERs of 72.00%, 57.94%, 62.22%, and 64.71% for the Gamo, Gofa, Dawuro, and Wolaita languages, respectively. These results could serve as a baseline for future research on speech processing using these corpora and for the development of ASR systems for any of these languages.

2. Ometo Languages

Ethiopia has more than 83 different languages with up to 200 different spoken dialects (Tonja et al., 2021). The Ethiopian languages are divided in to four major language family groups. These are Semitic, Cushitic, Omotic and Nilo-Saharan (Mara, 2018). The Omotic language classification history can be split into pre- and post-Fleming periods. "Omotic" did not exist prior to Fleming study on 1976; it was referred to as Southwest Cushitic, a subset of the Cushitic. Fleming, however, claimed that Omotic and Cushitic languages are essentially distinct from one another in a 1976 study he released that examined the lexical and grammatical characteristics of Omotic languages in contrast to Cushitic languages. In his study, he recognized eight grammatical differences between Omotic and Cushitic languages; on the basis of these, he projected that Omotic should be regarded as a distinct Afroasiatic family as contrasting to a Cushitic subfamily. Since then, Omotic has been broadly regarded as a distinct Afro-asiatic family (Wondimu, 2010).

Because of their similar phonology, syntax, and vocabulary, a wide collection of languages within Omotic are grouped together as a genetic unit, which is represented by Ometo as one sect (Amha, 2017). Ometo is one of the subfamilies of Omotic language family with its own classification of subgroups depending on its internal classification, more or less similar classifications, with minor differences (Woldemariam, 2014). The subgroup is divided into North, South, East, and West Ometo according to the study of Fleming. Languages referred to as the Wolaita dialect group, such as Wolaita, Gamo, Gofa, Dawuro, and Dorze, are included in the Northern group which are under our consideration (Wondimu, 2010).

The languages spoken in the Wolaitta, Gamo, Dawuro, and Gofa regions of southern Ethiopia, once referred to as the North Omo Zone, are the subject area of this study which are all considered to be part of

the North Ometo subgroup. All of these languages have a considerable number of native speakers. Wolaita language is estimated to have around 3.3 million native and dialective speakers though the approximation of the number of speakers is differ significantly because of no agreement is reached for where the boundary of the language is (Balcha, 2020) while the Gamo language, also known as Gamotstso doona or Gamotstso locally, is spoken natively by 1,070,626 people according to Central Statistical Authority (CSA). Also Dawuro language is spoken by more than 600,000 people (Amamo, 2017) while Gofa language is spoken by more than 400,000 peoples(Hirboro, 2015). These languages have different functions in Ethiopia. They are taught as a course up until university level and as a language of instruction in the lowest elementary school grades.

All four speech variants share a large number of vocabulary and grammatical characteristics, making them mutually comprehensible. The languages share more over 80% of cognates, with Gamo and Dawuro sharing slightly fewer cognates—79%—than the other languages (Woldemariam, 2014).

Similar to English, the Ometo language has 34 letters that are derived from suffix-based languages and is written using an alphabetic writing system that is an expanded form of the Latin script. Ometo languages follow SubjectObject-Verb (SOV) word order and Dawuro, Gamo, Gofa and Wolaita languages shares 79% of consonant and 100% vowels inventories (Tonja et al., 2021). Apart from having similar characteristics, Ometo languages also show a considerable amount of partial and identical vocabulary. Other than this, all four languages have cognates, and very few vocabulary words are unique to one Ometo language (Woldemariam, 2014).

2.1. Phonology

Ometo languages share significant amount of phonetic properties. They have almost identical inventory of consonants though there are some distinctions in the four languages' phonemic inventories in terms of both quantity and kind. They each have twenty-five phonemes, whereas Gamo has twenty-six consonant phonemes. The only unique consonant in Gamo is /dz/; all other Ometo languages are lacking. There is no symbol for this sound in the orthography created for the four dialects. The four dialects' phonemic inventory also reveal differences in the consonants /t' and /s'/. While the other three have /s'/, Wolaitta has /t'/. The cognates of the four dialects consistently contain the consonants /t'/ and /s'/(Woldemariam, 2014).

When comparing the phonemic inventories of the four dialects, Wolaita stands apart from the other three due to two reasons: first, it does not have the alveolar affricate consonant "ts", and second, it has a unique phoneme, /t'/, that is not found in any other dialect. Nonetheless, the /t'/ in Wolaita consistently matches the /s'/ in the other languages. On the other hand, cognates demonstrate that the "ts" in Dawuro, Gamo, and Gofa correlates to the geminated tt in Wolaita. Furthermore, Gamo has a distinct consonant ('dz') that isn't present in any of the other three. Cognates demonstrate that 'dz' in Gamo matches to z(z) elsewhere. Gamo, Dawuro, and Gofa share more phonological characteristics with each other than Wolaita, according to the phonological inventories displayed above. Because 'dz' is missing from the other two (Dawuro and Gofa) phonemic inventory, Gamo appears to be quite different from the other two(Woldemariam, 2005). The vocalic inventory is the same for all the four languages. All four of the languages contain the same five phonemic vowels: five short and five long(Woldemariam, 2014).

In terms of phone set, Wolaita language consists of around 62 phone set. These are A, AA, B, BB, C, CC, D, DD, E, EE, F, G, GG, H, I, II, J, JJ, K, KK, L, LL, M, MM, N, NN, O, OO, P, PP, Q, QQ, R, S, SS, T, TT, U, UU, V, W, X, XX, Y, YY, Z, ZZ, 7, 77, CH, CHCH, DH, DHDH, NH, NYNY, PH, PHPH, SH, SHSH, ZH, ZHZH, TSTS. Dawuro language consists of around 59 phone set. These are A, AA, B, BB, C, CC, D, DD, E, EE, F, G, GG, H, I, II, J, JJ, K, KK, L, LL, M, MM, N, NN, O, OO, P, PP, Q, QQ, R, S, SS, X, XX, T, TT, U, UU, W, Y, YY, Z, ZZ, 7, 77, TS, SH, SHSH, PH, PHPH, NH, DH, DHDH, CH, CHCH, TH. Gamo language consists of around 62 phone set. These are A, AA, B, BB, C, CC, D, DD, E, EE, F, G, GG, H, I, II, J, JJ, K, KK, L, LL, M, MM, N, NN, O, OO, P, PP, Q, QQ, R, S, SS, T, TT, U, UU, W, X, XX, Y, YY, Z, ZZ, 7, 77, CH, CHCH, DH, DHDH, PH, PHPH, SH, SHSH, ZH, TH, DZ, THTH, DZDZ. Gofa language consists of around 59 phone set. These are A, AA, B, BB, C, CC, D, DD, E, EE, F, G, GG, H, I, II, J, JJ, K, KK, L, LL, M, MM, N, NN, O, OO, P, PP, Q, QQ, R, S, SS, T, TT, U, UU, W, X, XX, Y, YY, Z, ZZ, 7, 77, CH, CHCH, DH, DHDH, PH, PHPH, SH, SHSH, ZH, TH, DZ.

2.2. The writing system of Ometo Languages and its problems

The alphabet used in the orthography designed for Wolaita, Gamo, Gofa, and Dawuro combines an expanded Latin script with the Alphabetic Writing System. It has distinct punctuation rules, sound-symbol mapping, and its own punctuation system. There are thirty-four letters in the spelling under the orthography of Ometo languages. There are 68 symbols in all the orthography: one capital and one small form for each letter. Certain orthographically significant consonants that the Latin letters do not represent are present in the four languages(Woldemariam, 2014).

2.3. Tone-accent of Ometo languages

Research indicates that the languages in consideration are acknowledged as tone-accent languages. In these languages, the meaning of a word can vary as its tone pattern changes (Fanta, 2010; Woldemariam, 2014). For example:

- góda (lord) : godá (wall)
- ?áwa (where) : ?awá (sun)
- bita (witchcraft) : bita (bewitch!)
- t‘uma (darkness) : t‘uma (become dark!)

2.4. Consonant Gemination of Ometo languages

Gemination occurs in Ometo languages when two consonants appear consecutively in a word. A geminated sound is considered strong, whereas a non-geminated (or singleton) sound is considered weak. For instance, almost all consonants except /w, r, f, h, nh/ undergo a simple gemination. For example:

- Ola (Throw away!) : Olla (pit)
- mataa (Near) : maattaa (Grass)

Furthermore, complex consonant gemination, which involves the sequential occurrence of consonant phonemes in the form of $C_1C_2C_3$ within a word, occurs in the Ometo languages (Hirboro, 2015).

3. Corpora Preparation

Speech corpora with matching transcriptions are necessary for the development of any deep learning-based ASR (Automatic Speech Recognition) system, as well as to minimize the complexity that comes with utilizing several separate frameworks. To my knowledge, no attempts have been made to create ASR systems for the Gamo, Gofa, and Dawuro languages, despite the fact that scholars have made multiple attempts to do so for the Wolaita language. The work known for the development of standard speech corpora for Wolaita language is the development of the Large Vocabulary Read Speech Corpora for Four Ethiopian Languages (Abate et al., 2020).

In this study, we have, therefore, tried to develop read-mode speech corpora for four Ometo languages: Gamo, Gofa, Dawuro and Wolaita.

3.1. Text Collection and Pre-processing

For each of these languages, we have first gathered and pre-processed a sizable text corpus from the Internet and other sources, in accordance with the read-mode speech corpus building approach though Ometo languages are low resourced languages which have no enough text data available for conducting the research. It was really difficult to find the digital data for low resourced Ometo languages because most of the data available in these languages are found in hardcopy format rather in softcopy format which is easy to collect

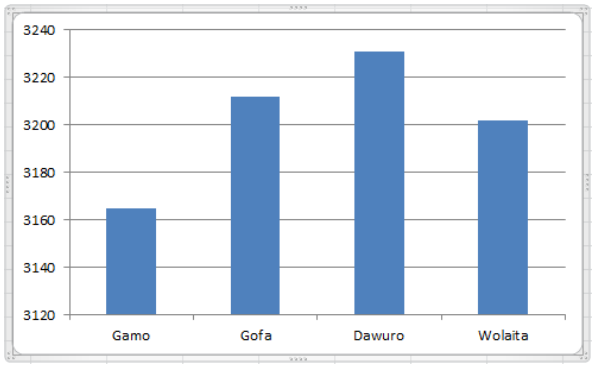

Figure 1: Textual data size of each language.

and preprocess to develop ASR systems. For this research, we have collected most of the text data from the freely available Ebible source [1]), which supports many other languages, and the rest from broadcasts, text books and other sources. In this way, we tried to incorporate textual information from other areas, although the spiritual area, which was mostly drawn from the ebb, accounted for a significant amount. After that, we pre-processed the gathered data to organize it in a manner that facilitates the creation of ASR systems. The following procedures were performed as part of this pre-processing:

- All words in the dataset were converted to lowercase to avoid ambiguity in word recognition.
- Extra spaces were removed because they could negatively impact the prediction accuracy of the ASR model.
- All numerical values were eliminated and substituted with their textual equivalents in order to improve system efficiency and guarantee uniformity in the text representation.
- All punctuation marks were eliminated. Even though the use of punctuation mark is necessary to determine the sentence boundary and avoid the ambiguity of meaning of the sentences, it is difficult to recognize them with limited vocabularies.

Then we have selected sentences using algorithm that uses phone set as units and produces phonetically rich and balanced corpus which encompasses all phone set according to their existence in the original data. In order to reduce reading difficulties, we have taken into consideration sentences that are less than 15 words. This also ensured the appearance of all phone set according to their presence in the original data.

Though Ometo languages suffer due to this scarcity, the corpus we have collected 12,810 preprocessed sentences for developing any ASR systems. To obtain these much text data 3165 sentences for Gamo, 3212 sentences from Gofa, 3231 sentences form Dawuro and 3202 sentences from Wolaita were collected. The contribution of each language is represented by figure 1 in above: Within the sentences phone sets that were listed above and existing in each language were represented in accordance of phonetically balanced and rich manner. The consecutive figures below demonstrates the appearance of every phone set in their original and finally selected form for four languages. The phone set with less frequency are taken fully from the original dataset to the selected one in order to avoid possible disappearance during the training of any ASR model for four languages.

Aiming at having at least 100 clean audio files from each speaker, we have assigned 130 utterances for each of the 25 Gamo, Gofa, Dawuro and Wolaita speakers(Abate et al., 2020). After intial recording, we have identified unread, corrupted and unclean data and recording again using 4 speakers for Gamo language, 2 speakers for Gofa language, 1 speaker for Dawuro language, and 3 speakers for Wolaita language. Distinct set

1. http://ebible.org

of prompts for each speaker in each language have been randomly selected from the prepared text databases. A total of 110 speakers participated in the recording process across four languages.

3.2. Speech Recording and Preprocessing

We used two TECNO POVA Neo smartphones and one Samsung smartphone to record the speech. Lig-Aikuma, an Android-based speech recording app, was installed on all three phones for recording (Gauthier et al., 2016). The purpose of this open-source smartphone application is to document endangered languages in linguistic fieldwork. Its primary objective is to make the recording, transcription, and translation of oral language data easier. It was created by the Laboratoire d'Informatique de Grenoble (LIG). The software is configured to record waves from a single channel using 16kHz sampling and 16bit pulse-code modulation (PCM) encoding. We have used a semi-supervised recording technique with this configuration. The operation of the Lik-aikuma recording system was briefly explained to the speakers who were then free to select a time and location that worked best for them. Additional care was taken to record in a quiet setting. As they completed reading, the audio file was saved. The recording program displayed the text one sentence at a time. The recordings were conducted at Mirab Abaya Kebele for the Gamo language; Sawula Campus and Wolaita Evangelical Seminary (WES) for the Gofa language; and Dawuro Tarcha Campus for the Dawuro language. Wolaita speech data was collected from various locations, including a preparatory school, private colleges, and Wolaita Sodo University.

The readers for the Dawuro language were selected from university students in the Dawuro Language Department, while those for the Gofa language were mainly selected from students in the Gofa Language Department at Sawula Campus. For Gamo language, the readers were selective from teachers in Mirab abaya secondary school and also others were included with checking their ability to read Gamo language outside the school. For the Wolaita language, readers were selected based on their proficiency, including students from the Wolaita Language Department at Wolaita Sodo University and Wolaita language teachers from various secondary schools across Sodo and Boditi town. In selecting the readers native speakers are considered. The whole recording process took place for about 2 months.

After completion of recording process, we have filtered audio data. Some of them had only a few sentences, some were empty files, and some were just word or phrase repetitions that failed the text to speech alignment stage. Furthermore, we have conducted some preprocessing step in order to enhance the quality of the speech corpora.

- Normalization: By altering the audio signal's amplitude levels, the ASR system's accuracy and dependability are improved and constant volume levels are maintained throughout recordings.
- Noise reduction: It aims to improve the quality of speech signals by reducing undesired background noise.

Then, we have aligned processed speech data to its corresponding transcription using python code in the manner suitable for training any ASR system.

3.3. Details of the Speech Corpora

All Automatic Speech Recognition (ASR) systems rely heavily on high-quality speech data for training. In order to train any Automatic Speech Recognition (ASR) system effectively, speech data collecting is a painstaking procedure that gathers a variety of high quality audio recordings. Determining the goals and parameters of the data collection, along with the target language, domain, and any potential differences like accents or dialects, is an essential initial step. To guarantee representation across demographic variables such as age, gender, and language background, we have chosen participants carefully as much as possible. Though qualifying the criteria in full extent to have better corpora seems to be difficult, we tried to go along with the standard to include different age groups ranging from 18 up to 50, choosing for environments with minimum background and balancing between genders for the aim of any Automatic Speech Recognition model to be tolerant with different age groups and genders.

According to a detailed examination of the usable corpora, recordings of 29 Gamo, 27 Gofa, 26 Dawuro, and 28 Wolaita speakers have been made. However, not every speaker has read all 130 sentences. Therefore, we represent the distribution of the number of audio data per speaker, age and gender distributions for each language in Table 1, 2 and 3.

Table 1: Number of Utterances per speaker

No of utter.	Gamo	Gofa	Dawuro	Wolaita
90 - 99	5	3	2	4
100 - 109	9	11	13	9
110 - 119	7	6	6	8
120 - 130	8	7	5	7
Total	29	27	26	28

Table 2: Age Distribution of the speakers

Age range	Gamo	Gofa	Dawuro	Wolaita
18 - 35	20	12	13	16
36 - 50	9	15	13	12
Total	29	27	26	28

Table 3: Gender Distribution of the speakers

Gender	Gamo	Gofa	Dawuro	Wolaita
Male	16	18	14	14
Female	13	9	12	14
Total	29	27	26	28

4. Development of Baseline ASRSs for Four Ometo Languages

Creating ASR systems for each of the four languages allowed us to examine how well the stated corpora could be used to develop speech recognition model. Exception to Wolaita language, which has already been done, we consider our findings as baselines for future utilization of the corpora in the development of ASR systems for these languages (Abate et al., 2020). The details are presented in the following subtopics.

4.1. System Architecture

Our system architecture starts with loading speech data with its corresponding transcription. Then, whole corpora split into training and testing along with pre-processing tasks as we have discussed in Subsections 3.1 and 3.2. This step was followed by feature extraction using Short Time Fourier Transform (STFT) by assigning frame length, frame step and fast fourier transform (fft) length. The audio stream is divided into brief, overlapping windows using the Short-Time Fourier Transform (STFT), and the Fourier Transform is applied to each window to determine its frequency content over time. As a result, a spectrogram is produced, which represents each frequency component's intensity as a function of time and captures the spectral and temporal information that are essential for speech recognition. A feature extraction procedure was needed to generate spectrograms, which were later fed into a CNN for further analysis and acoustic modeling.

High level spatial features are extracted from the image using CNN (Bhatta et al., 2020). Since the STFT plot can be viewed as a time-varying frequency intensity that resembles an image, a two-layer 2D CNN was used to extract high-level spatial features. Two 2D CNN layers were utilized: a first layer with filters of 32,

kernel size of [11, 41], strides of [2, 2], and a second layer with filters of 32, kernel size of [11, 21], strides of [1, 2]. CNN can obtain more robust features by utilizing maximum pooling and local filtering strategies in comparison to conventional speech features.

The sequential data was learned using a Gated Recurrent Unit (GRU). The GRU, which consists of two gates—the reset gate and the update gate—regulates the flow of information to address the vanishing gradient problem commonly found in RNNs. The reset gate is in charge of figuring out how much historical data should be erased. The update gate is in charge of figuring out how much historical data should be sent forward for future use (Bhatta et al., 2020). The training process is greatly streamlined by CTC's ability to directly map an input speech sequence into a string of text sequences, maximizing the likelihood of both the input and output sequences. The core of the CTC-based acoustic model is still a sequence classification issue, where each neural network node's output chooses the generation path with the highest probability. As a result, the relationship between the CTC's input and output is frequently many-to-one. The auditory feature parameters are subsequently retrieved by means of a convolutional neural network, followed by seven layers of Bi-directional GRU and the SoftMax layer, once the CTC-based acoustic model has recognized speech. Thus, the output sequence is each node's maximum probability label. Ultimately, the recognition outcome is produced by the CTC decoding algorithm's optimized output label sequence (Sung et al., 2023). Finally, Multilingual recognition model which was output of training of the model with training dataset evaluated using WER to determine the performance of the developed model.

4.2. Experimental setups

To train and test our multilingual recognition model in deep learning, we have used different hyper-parameters. The table 4 below shows different hyper-parameters with their suitable values.

Table 4: Selection of Hyper-parameters of the model

Hyper-parameter used in the research	
Optimizer	Adam
Learning rate	1e-4
Activation function	Softmax
Learning rate	1e-4
GRU gate activation function	Sigmoid
GRU activation function	Tanh
Dropout	0.5
RNN type	Bi-GRU
Epoch	60
Bi-direction GRU layer	7
CNN activation function	ReLU
CNN layer	2
RNN units	128

4.3. Experiment result

Both Training and testing speech corpora divided in the manner of including both gender to represent in balanced way in the development phase of the ASR model. For Gamo language, using speech corpora of 5.5350 hrs with split ratio 5.01 hrs for training and 0.5250 hr for testing, we have obtained WER of 72.00%. For Gofa language, using speech corpora of 5.97 hrs with split ratio 5.4386 hrs for training and 0.5314 hr for testing, we have obtained WER of 57.94%. For Dawuro language, using speech corpora of 6.8103 hrs with split ratio 6.1453 hrs for training and 0.665 hr for testing, we have obtained WER of 62.22%. For Wolaita language, using speech corpora of 6.0358 hrs with split ratio 5.4184 hrs for training and 0.665 hr for testing, we have obtained WER of 64.71%. Table 5 below shows their respective WER result.

Table 5: WER result

Languages	WER result
Gamo	72%
Gofa	57.94%
Dawuro	62.22%
Wolaita	64.71%

5. Conclusion and Recommendation

The development of four speech corpora for the four Ometo languages—Gamo, Gofa, Dawuro, and Wolaita—is presented in this study. A total of 24.3511 hours of voice recordings were collected using all four languages. The performance of the baseline ASR systems, which were developed using the corresponding corpora, has also been reported. The WERs we obtained demonstrated that the corpora are appropriate for more research and the creation of ASR systems suited to these languages. The information will be disseminated to the scientific community in order to promote the advancement of speech processing components.

Finally, we would like to suggest some points as recommendation for further study.

- Better result would be resulted by multiplying the amount of the dataset used for training multilingual recognition model.
- Using mobile phone with high quality of sound recording would increase the quality of data and in turn increase the quality of the model developed.
- Considered dialects for model building and decoding.
- Appending the LM independently to the model.

Acknowledgments

For their financial assistance with the data collection process, Wolaita Sodo University has our deepest gratitude. Additionally, We want to express our gratitude to mentors for their unwavering support and helpful criticism during this research. Their encouragement and assistance were crucial to finish our research.

References

Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Afnafu, and Binyam Ephrem Seyoum. Large vocabulary read speech corpora for four ethiopian languages: Amharic, tigrigna, oromo and wolaytta. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4167–4171, 2020.

Alemayehu Asfaw Amamo. Enset value chain, the case of dawuro zone, southern nations nationalities and peoples regional state, ethiopia. 2017.

Azeb Amha. The omotic language family. Cambridge University Press, 2017.

Tekalgn Balcha. A thesis submitted to department of information technology, school of graduate studies, wolaita sodo university. 2020.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014.

Bharat Bhatta, Basanta Joshi, and Ram Krishna Maharjhan. Nepali speech recognition using cnn, gru and ctc. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 238–246, 2020.

H Fanta. Speaker dependent speech recognition for wolaita language. *Addis Abeba University, Msc Thesis*, 2010.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. Lig-aikuma: A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*, 2016.

Sellassie Cheru Hirboro. Documentation and grammatical description of gofa. *Addis Ababa Univeristy*, 2015.

Sendong Liang and W Yan. Multilingual speech recognition based on the end-to-end framework. *Multimedia Tools and Applications*, 2022.

Melaku Mara. *English-Wolaytta Machine Translation using Statistical Approach*. PhD thesis, St. Mary's University, 2018.

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

Markus Müller and Alex Waibel. Using language adaptive deep neural networks for improved multilingual speech recognition. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 167–172, 2015.

Suman K Saksamudre, PP Shrishrimal, and RR Deshmukh. A review on different approaches for speech recognition system. *International Journal of Computer Applications*, 115(22), 2015.

Wen-Tsai Sung, Hao-Wei Kang, and Sung-Jung Hsiao. Speech recognition via ctc-cnn model. *Computers, Materials & Continua*, 76(3), 2023.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. Dnn-based multilingual automatic speech recognition for wolaytta using oromo speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 265–270, 2020.

Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, Mesay Gemeda Yigezu, and Ethiopia Hossaena. Low resourced multilingual neural machine translation for ometo-english. *Gamo*, 125(23,589):16, 2021.

Hirut Woldemariam. Notes on the north ometo dialects: Mutual intelligibility tests and structural variation. *Cushitic–Omotic Studies*, pages 79–112, 2005.

Hirut Woldemariam. Writing both difference and similarity: towards a more unifying and adequate orthography for the newly written languages of ethiopia: the case of wolaitta, gamo, gofa, dawuro. *Journal of Languages and Culture*, 5(3):44–53, 2014.

Henok Wondimu. The grammaticalization of copula markers in the ometo subgroup. *Addis Ababa University*, 2010.