AUTOMATIC SPEECH RECOGNATION FOR AMHARIC LANGUAGE USING SELF-SUPERVISED LEARNING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic Speech Recognition (ASR) systems have become a very natural human-machine interaction in which it allows users to speak entries rather than punching numbers on a keypad. This study presents an automatic speech recognition system for Amharic language that is one of the Ethiopian Languages. The current attempts on speech recognition systems require thousands of hours of transcribed speech dataset to reach adequate accuracy. However, the large majority of under-resourced spoken languages in general, Ethiopic languages in particular, have a very limited amount of labeled speech dataset. Thus, in this study, a self-supervised Transformer based Wave2Vec 2.0 approach has been conducted to build an ASR system for Amharic language. In order to pretrain the proposed model in an unsupervised approach, a total of more than 200 hours of noisy and clean unlabeled speech data from different local media have been collected. In addition, a total of 30 minutes of labeled speech dataset that contains speech utterances and the corresponding transcription text has been prepared. Then, a Wave2Vec model that is initialized with weighted parameters has been pre-trained on the unlabeled speech dataset to construct the speech vector representation. Then the pretrained model has been fine-tuned using a small amount of labeled speech dataset. A standard Word Error Rate (WER) evaluation metric has been used to evaluate the fine-tuned ASR model and it has shown a clear significant result.

1 INTRODUCTION

Speech is one of the prevalent forms of expressing oneself. A person can speak 5000 – 7000 words every day. Unlike other natural language processing tasks, speech synthesis, language identification, speech summarization and translation problems have gained attention very late in the period of human language technology, and it is now an active research topic. Automatic speech recognition (ASR) is the task of recognizing human speech and converting it to text Papastratis (2014). ASR technology is one of the important technologies for all human languages to directly process speech in human-machine interaction. As a result, a large number of researches and automatic speech recognition systems in a variety of human languages have already been developed (Vu et al. (2014); Li et al. (2019); D N et al. (2021); Teferra et al. (2020)). However, only a few percent of the languages have taken into account from the 7000 languages in the world. The main reason is, there is no a large enough speech data for each language. Speech recognition models require large amount of transcribed speech data to achieve acceptable performance. We lack such labeled data for a large number of human languages. Almost all Ethiopian languages are under-resourced, and they are among those that are not benefiting from the advancement of spoken language technologies Yifiru et al. (2020). Even though many languages are spoken in Ethiopia, Amharic is the dominant language that is spoken as a mother tongue by a large segment of the population and it is the most commonly learned second language throughout the country. It is a working language of the government of Ethiopia as well. Amharic has its own writing system that is called "Ge'ez script". It is the second largest spoken Semitic language in the world next to Arabic RACOMA (2013). As a result, having an Amharic speech recognizer is useful in various aspects of life. These ASR applications include dictation systems, command and control systems, telephony systems, meeting transcription

systems, information retrieval systems, broadcast news transcription systems, and computer aided instruction systems Teferra et al. (2015).

Collecting a large amount of labeled speech data is very expensive and time-consuming. In addition, labeled data is significantly more difficult to get in many contexts than unlabeled data. Furthermore, the majority of existing ASR systems employ data by forced segmentation alignment and multi-module training as acoustic, pronunciation, and language modeling techniques. Most recently, self-supervised learning has gained much attention, and demonstrated to work well both in low and high-resourced labeled data settings for ASR. In this paper, we apply the wav2vec 2.0 framework that attempts to build an accurate speech recognition model with a small amount of transcribed data. It is a self-supervised neural network because it is pre trained only on unlabeled data. In this work, the Amharic unlabeled audio data has pre-trained using the wav2vec model for weight initialization, then fine-tuned the pre-trained model with a small amount of Amharic labeled speech dataset. Word error rate (WER) has been used to evaluate the model. WER calculates the number of words that are incorrectly transcribed by the model and compares with the correct data version to evaluate the model.

2 Related Work

Research in ASR for Amharic began in early 2001 with the work of Berhanu (2001) who developed an isolated Consonant-Vowel syllable recognition system using the Hidden-Markov Modeling Toolkit. As the researcher has noted, the result seems to be low in comparison to other systems that have developed for other languages. This could be due to issues with the recording environment or a lack of training speech data. Consequently, several research studies have been conducted in the area of Amharic automatic speech recognition (Tadesse (2002); Yifiru (2003); Girmaw (2004); Tefera (2005)) using the Hidden-Markov Modeling Toolkit and (Teferra et al., 2020) using deep neural network. These researchers have mentioned in their studies that there is a limitation of labeled speech corpora.

D N et al. (2021) has proposed wav2vec2.0 models for the speech recognition task for three Indian languages that are Telugu, Tamil, and Gujarati. They have shown that fine-tuning with less than 10 hours can give competitive performance compared to the previous state-of-the-art supervised method. Baevski et al. (2020) has also proposed a wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations in their study. They have trained the model using only ten minutes of labeled data and pre-training on 53 000 hours of unlabeled data and achieved 4.8 WER. In their work they have clearly shown the feasibility of speech recognition with limited amounts of labeled data. Self-supervised learning has emerged as a paradigm in machine learning for learning general data representations from unlabeled examples and fine-tuning the model on labeled data. Therefore, recent works have revealed that robust sequential models with a deep network architecture promise to support very low resourced language model training with minimal effort of human intensive labeled dataset preparation activities. Thus, in this study we have proposed a self-supervised Transformer based Wave2Vec 2.0 model for Amharic speech recognition system.

3 Methodolgoy

In this work more than 500 hours of unlabled audio data have been collected. This data incorporates multiple speakers, genders and age groups, and also a variety of domain sources such as broadcast news, broadcast conversation and TV dramas. The data has been cleaned from noise and prepared for model training. The preprocessing steps that have been taken for the unlabled data prapration are data collection, sampling and segmentation. Then, 200 hours of data have been extracted from the total data after preprocessing. Each of the segmented audio files have been sampled at 16 kHz with 16-bit resolution and saved in the *.wav format.

Labled data has been part of the model training. In this study the steps that have been taken to make the labeled data ready for the speech recognition system are the following: Data collection, sampling, segmentation, lexicon and dictionary preparation and language model building. The first three data preparation steps that are the data collection, sampling and segmentation have been done on the unlabeled data preparation as well. Like the unlabled data, each of the segmented audio files have been sampled at 16 kHz with 16-bit resolution and saved in the *.wav format. During the

lexicon and dictionary corpus preparation, data cleaning such as spelling and grammar have been done, abbreviations have been expanded, and numbers have been transcribed. Finally, a total of 30 minutes of labeled speech dataset has been prepared. The dataset contains pairs of spoken utterances and their associated transcripts.

For the experiments, the publicly available wav2vec 2.0 model has been used. First, the English wav2vec model has been used for weight initialization and pre-trained a self-supervised model on the 200 hours of Amharic unlabeled speech dataset. Then, the pretrained model has been fine-tuned using the 30 minutes of labeled speech dataset. After that, a language model and lexicon file have been built using the labeled data and the text corpus. Finally, the three models independently and the pre-trained model togehterwith fine tune model and language model have been evaluated using word error rate (WER).

4 RESULT

As the models evaluation result shows, overall the models performance has been evaluated as a system. The study has been demonstrated that speech recognition system development is possible with very low-resource using self-supervised learning on unlabeled data. The pre-trained model has been evaluated with only 30 minutes fine-tuned labeled dataset, and achieved 35.6 WER. The word error rate is higher for longer sentences than for short sentences. Still, we only have trained the model on 200 hours of unlabeled data and only 30 minutes of labeled data for the fine-tuned model. This shows that the results are promising.

5 CONCLUSION

In this paper, we have presented an Automatic speech recognition system that has performed for the Amharic language. The system has been developed using a self-supervised transformer based Wave2Vec 2.0 model approach. Although developing speech recognition systems for low-resourced languages from scratch is very challenging in speech processing. This study has shown that we can build a speech recognition model with promising performance using a small amount of labeled data. Hence, this research paper is ongoing research, the model performance will be improved by extending the labeled and unlabeled speech corpus.

REFERENCES

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. 2020.

Solomon Berhanu. Isolated amharic consonant-vowel syllable recognition: An experiment using the hidden markov model. 2001.

K. D N, P. Wang, and B. Bozza. Using large self-supervised models for low-resource speech recognition. 2021.

M. Girmaw. Isolated amharic consonant-vowel syllable recognition: An experiment using the hidden markov model. 2004.

X. Li, S. Dalmia, A. W. Black, and F. Metze. Multilingual speech recognition with corpus relatedness sampling. 2019.

I. Papastratis. Speech recognition: a review of the different deep learning approaches. 2014.

B. E. R. N. A. D. I. N. E. RACOMA. Amharic: Ethiopia's official language. 2013.

K. Tadesse. Sub-word based amharic speech recognizer: An experiment using hidden markov model (hmm). 2002.

S. Tefera. An amharic speech corpus for large vocabulary continuous speech recognition. 2005.

S. Teferra, M. Yifiru, and W. Menzel1. Amharic speech recognition: Past, present and future. 2015.

S. Teferra, M. Yifiru, and T. Schultz. Deep neural networks based automatic speech recognition for four ethiopian languages. 2020.

T. Vu, D. Povey, D. Imseng, and P. Motlicek. Multilingual deep neural network based acoustic modeling for rapid language adaptation. pp. 7639–7643, 2014.

M. Yifiru. Automatic amharic speech recognition system to command and control computers. 2003.

M. Yifiru, S. Teferra, and T. Schultz. Dnn-based multilingual automatic speech recognition for wolaytta using oromo speech. pp. 265–270, 2020.