Wav2Vec-Based Self-Supervised Learning for Court Hearing Transcription

Anonymous ACL submission

Abstract

An automatic court hearing transcription system is being developed for the Federal Supreme Court of Ethiopia to address the challenges faced in manual transcription. By utilizing Automatic Speech Recognition technology, the system aims to transcribe Amharic language court recordings accurately and efficiently. This innovative solution not only improves the court system but also safeguards the health of transcribers and enhances the overall speed and quality of legal proceedings in Ethiopia. In this study, a self-supervised Transformer based Wave2Vec 2.0 approach has been conducted to build an ASR system. With a dataset comprising over 500 hours of unlabeled data, the system has achieved a remarkable Word Error Rate (WER) of 14.36%, showcasing its effectiveness in transcribing court proceedings with high accuracy.

1 Introduction

The development of speech processing applications for less-resourced languages faces significant challenges due to the lack of language and technological resources. In Ethiopia, where over 80 less-resourced languages exist, efforts have been made to bridge this gap. Researchers and institutions have been preparing speech corpora for Ethiopian languages, including Amharic, Afan Oromo, Tigrigna, and Wolaytta[1]. These efforts have primarily focused on broadcast media and religious domains, with limited attention to domain specific resources such as the judicial domain.

In the legal sector, court hearing recordings in Ethiopia are transcribed manually, leading to time-consuming processes and potential errors. Manual transcription poses health risks to transcribers and results in unnecessary delays in court proceedings. To address these challenges, Automatic Speech Recognition engines have emerged as a frontline solution for accurate and efficient legal transcription. This paper presents the development of an Amharic court hearing transcription system, aiming to leverage ASR technology to streamline transcription processes in the legal sector.

The system not only fills a crucial gap in exploring the potential of automatic speech processing applications for less-resourced languages but also contributes to the creation of an in-domain speech corpus for the Amharic language. By focusing on domain specific speech resources, this project seeks to enhance the efficiency and accuracy of legal transcription in Ethiopia while paving the way for future advancements in speech technology for societal benefit.

2 Related works

Research in Automatic Speech Recognition for Ethiopian languages, like Amharic, has focused on command-and-control systems and real time transcription. Early projects, such as [2], achieved high accuracy using limited vocabulary and Hidden Markov Models (HMM). In the judicial domain, systems like "Brana" [3] faced challenges with spontaneous speech recognition, resulting in varying Word Error Rates (WERs).

Efforts funded by the Ministry of Communication and Information Technology (MCIT) produced dictation systems with moderate WERs [4]. Internationally, ASR systems for judicial use in languages like Polish, Italian, and Slovak showed WERs ranging from 5.26% to 50% [5][6]. Recent improvements in UK Supreme Court transcription reduced WERs to 11.6% [7]. Existing ASR systems for Ethiopian languages often rely on older methods and small datasets. 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 selfsupervised 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.

3 Methodology

In this study, a total of 500 hours of Amharic unlabeled speech data was collected from the Federal Courts of Ethiopia, along with 62 hours of labeled speech datasets containing paired spoken utterances and their transcripts. The data preparation steps for the speech recognition system included data collection, sampling, segmentation, lexicon and dictionary preparation, and language model building.

After collecting the speech corpus from the Federal Courts of Ethiopia, the next step involved sampling and segmenting the audio files at 16 kHz with 16-bit resolution, saved in the *.wav format. The text corpus underwent corrections for spelling and grammar errors, expansion of abbreviations, and transcription of numbers.

For the experiments, the publicly available wav2vec 2.0 model was used. Initially, the English wav2vec model was employed for weight initialization,followed by pre training a self-supervised model on the 500 hours of Amharic unlabeled speech dataset. Subsequently, the pre-trained model was fine-tuned using the 62 hours of labeled speech datasets. A language model and lexicon file were built using the labeled data and text corpus. Finally, the pre-trained model and fine tuned model were evaluated with the language model.

4 Result

The Automatic Speech Recognition (ASR) system for court hearing transcription was evaluated using a 15.93-hour evaluation set, representing 10% of manually segmented and transcribed speech data. The Word Error Rate (WER) of the system compared to the test transcription was found to be 35.3%. However, a detailed error analysis revealed that many errors were present in the reference transcription, impacting the WER rate. When evaluated by experienced transcribers listening to 6,000 utterances and cross-checking with the ASR system output, the WER of the system improved to 14.36%, indicating an accuracy of 85.64%.

5 Conclusion and Future Work

The development of an Automatic Speech Recognition (ASR) system for Amharic court hearing transcription demonstrates significant advancements in addressing transcription challenges in the Ethiopian legal sector. By leveraging ASR technology and utilizing a substantial in-domain speech corpus, the system achieved a notable Word Error Rate (WER) of 14.36%, signifying an accuracy of 85.64%. This innovative solution enhances the efficiency and accuracy of legal transcriptions while mitigating health risks and delays associated with manual processes. Future work will focus on refining the ASR system by expanding the dataset with more diverse speech samples, incorporating advanced machine learning algorithms, and exploring cross-lingual transfer learning techniques for other Ethiopian languages. Additionally, integrating the ASR system with real-time court proceedings and developing user-friendly interfaces for practical deployment will be prioritized to ensure broader adoption and societal impact.

References

1. *Abate, S.T., Tachbelie, M.Y., Melese, M., Abera, H., Abebe, T., Mulugeta, W., Assabie, Y., Meshesha, M., Afnafu, S., Seyoum, B.E.: Large vocabulary read speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4167–4171. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.513*

2. *Tachbelie, M.Y.: Application of Amharic*

Speech Recognition System to Command and Control Computer: an Experiment with Microsoft Word. Ph.D. thesis, School of Information Studies for Africa, Addis Ababa University (2003)

3. *Alemayehu,B.A.,Abate,S.T.: Brana(): application of amharic speech recognition system for dictation in judicial domain. In: In the proceedings of Ethiopian Information Technology Anual Conference (2015)*

4. *Tachbelie, M.Y., Abate, S.T., Abebe, E.: Development of morpheme based dictation system with the support of automatically constructed dictionary. Tech. rep., School of Information Science, Addis Ababa University (2015)*

5. *Lööf, J., Falavigna, D., Schlüter, R., Giuliani, D., Gretter, R., Ney, H.: Evaluation of automatic transcription systems for the judicial domain. In: 2010 IEEE Spoken Language Technology Workshop. pp. 206 211(2010). https://doi.org/10.1109/SLT.2010.5700852*

6. *Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Sabo, R., Pleva, M., Ritomský, M., Lojka, M.: Slovak automatic dictation system for judicial domain. In: Vetulani, Z., Mariani, J. (eds.) Human Language Technology Challenges for Computer Science and Linguistics. pp. 16–27. Springer International Publishing, Cham (2014)*

7. *Saadany, H., Breslin, C., Orăsan, C., Walker, S.: Better transcription of uk supreme court hearings (2)*