Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/158/2025.TJ23214

Study for Automatic Speech Recognition for Wav2Vec2.0

Xiwei Huang

Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
xiwei.huang@mail.utoronto.ca

Abstract: Automatic Speech Recognition (ASR) is a popular technology that converts speech audio into corresponding text. This application serves critical roles in areas such as virtual assistants, transcription services, accessibility tools, etc. This paper mainly introduces the application of the Wav2Vec2.0 model, which is an advanced self-supervised ASR model. The dataset used in this research is the Mozilla Common Voice dataset, which contains audio data in multiple languages and from people across different ages, genders, and occupations. In addition, the data preprocessing process and the architecture of the model will also be discussed in this research. The implementation demonstrates the strong ability of the Wav2Vec2.0 model in transcribing speech data from the Mozilla Common Voice database, and the experimental results highlight the model's robustness in handling variations in accent, speaking speed, and recording quality, achieving competitive word error rates (WER) across diverse linguistic scenarios. Results also indicate potential improvements in accuracy through more careful and targeted data processing and improving the tokenizer. All these findings underscore the model's future capacity in real-world speech recognition systems, emphasizing its adaptability and efficiency.

Keywords: Automatic Speech Recognition(ASR), Wav2Vec2.0, Common Voice, LibriSpeech

1. Introduction

Automatic Speech Recognition (ASR) is a hot field lately, and it is playing a key component of many applications ranging from voice-controlled assistants to accessibility tools for people with disabilities. ASR, which can be defined as a technology that transforms audio content into corresponding text transcription, is an important methodology in man-to-machine interaction. This increase in ASR use can be found in current applications such as Apple's Siri, Amazon's Alexa, Google Home, Microsoft's Cortana, or NICT's VoiceTra, which demonstrates the importance of ASR in an efficient and smart user interface [1]. With the expanding growth of audio and speech data, the consideration for accurate and efficient transcriptions from speech to text becomes more important. This is useful not only for improving user interaction but also for data-driven processes and content accessibility.

In addition, ASR sometimes merges with technologies from other fields, like natural language understanding (NLU) and machine learning. This merging further expanded its capacity and enabled it to have more meaningful interactions. Users will also get more personalized experiences through this cross-field merging. For example, there are ASR models with sentiment analysis and emotion recognition. The benefit of this combination is to better understand the user's intention and the whole context, which makes it act in a more intuitive and human-like mode.

DOI: 10.54254/2755-2721/158/2025.TJ23214

The tasks and application scenarios for ASR have changed over the years from simple to complex conditions, for example, from small to large-sized vocabulary, from speaker-dependent to speaker-independent, from read speech to spontaneous speech, and from quiet office rooms to noisy environments. In the earliest time, the 1950s-70s, people were using the dynamic time warping method to solve ASR questions. Then, moving to the 1980s- 90s, the hidden Markov model, known as HMM, became available and acted as a base approach at that time. Coming back to current days and the recent breakthroughs, Microsoft reports a Word Error Rate(WER) of 5.1% on the Switchboard corpus. These results are even better than real human performance, which underscores the power of deep learning with this remarkable progress [1]. According to a more recent study, current ASR research from 2015 to 2020 has shifted its focus to address some key performance influence factors, such as dialect variations, background noise, and some interruptions during speech [2]. Deep Neural Network (DNN) techniques and audio-visual integration methods have been utilized to enhance the robustness of the model in the situation described above. In neural machine translation and natural language processing (NLP), transformer-based models have been widely used and maturely applied. These transformer-based models have been increasingly explored for ASR, demonstrating promising performance improvements by efficiently capturing long-range dependencies within speech sequences [3]. Besides these advancements, there are still some major challenges that remain: handling background noise, simultaneous conversations(the “cocktail party” problem), microphone quality issues, vocabulary size, dialect diversity, and spontaneous speech pronunciation variations [2]. Solving these challenges would be influential for achieving high ASR accuracy in real-world applications.

In this study, the Wav2Vec2.0 model will be implemented and evaluated using both the Mozilla Common Voice and Librispeech datasets. The primary objective is to assess the model's effectiveness in diverse speech conditions and explore methods to enhance recognition accuracy through preprocessing and data handling techniques.

2. Implementation

In this research, the Wav2Vec2.0 model, introduced in 2020, is implemented and fine-tuned on both the Mozilla Common Voice and Librispeech datasets. The primary goal is to evaluate its effectiveness on speech recognition tasks using datasets with different characteristics. The implementation utilized Hugging Face’s transformer library, known for its efficient use of pre-trained models, fine-tuning capabilities, and integrated evaluation metrics. The core components discussed in this section include data loading, preprocessing, audio augmentation, training configurations, and metric evaluations.

For data loading and preprocessing, a custom PyTorch dataset class is designed to load audio files in MP3 or FLAC format. Each audio file was loaded using the torchaudio library, resampled to a 16kHz sampling rate, and converted to mono to maintain consistency across all samples. To enhance model robustness, audio augmentation is conducted during training. To simulate real-world variations, random speed perturbation, silence insertions, and gain adjustments are introduced. This significantly diversifies the training data and improves generalization. In particular, speed perturbation has been shown to be an effective and simple method for improving ASR performance, especially in low-resource scenarios [4].

A custom data collator integrates with the Wav2Vec2.0 processor, handling dynamic padding and tokenization of audio and transcripts. The transcripts are preprocessed to ensure compatibility with the tokenizer, including conversion to uppercase and removal of unwanted characters. The dataset is split into train, validation, and test sets with no duplicate data across them.

Training uses the Hugging Face Trainer class with specified hyperparameters. A batch size of 4 and gradient accumulation of 4 result in an effective batch size of 16. A learning rate of 3e-5 was chosen based on the original Wav2Vec2.0 paper [5]. Gradient checkpointing and half-precision are

DOI: 10.54254/2755-2721/158/2025.TJ23214

enabled to optimize resources. Early stopping is managed by selecting the best model based on the WER metric after each epoch. Both WER and Character Error Rate (CER) are calculated for every validation set. A custom function decodes model predictions and labels them into text before calculating WER or CER. Then, it will utilize either Hugging Face's evaluate library or Jiwer and integrate directly within the Trainer class to ensure flexibility and reliability.

After training, post-analysis, including logging and visualization of loss and WER/CER per epoch, will be reviewed and analyzed. The model's performance is assessed, and areas for potential improvement are identified. The implementation is structured to be reproducible and modular, clearly separating data preparation, training, and evaluation phases. Models and processors are saved systematically for future use.

2.1. Database

2.1.1. Mozilla Common Voice

The Common Voice is known for its commitment to openness and multilingualism. As described by, the project prioritizes scale and inclusiveness by combining community-based data collection and validation processes with open licensing (CC0) [6]. The authors note that over 50,000 contributors had participated in creating more than 2,500 hours of validated audio at the time of publication, underscoring the dataset's value for low-resource and multilingual ASR research. In this research, version 20.0 of the Common Voice Delta Segment was used. This dataset contains 1015 voices with a total duration of up to 45 hours, and the recordings are collected through a crowdsourced framework [7].

Each recording is provided with its corresponding transcript and includes optional speaker demographic information. All audio is released in MPEG-3 format at a 48kHz sampling rate and is resampled to 16kHz during preprocessing [6]. The English subset of the dataset was selected to align with the model's pretraining domain.

2.1.2. LibriSpeech

The LibriSpeech corpus is a widely-used ASR benchmark obtained from English audiobooks in the LibriVox project. "This paper presents the LibriSpeech corpus, which is a read speech data set based on LibriVox's audio books. The corpus is freely available under the very permissive CC BY 4.0 license" [8]. The dataset includes approximately 1000 hours of read English speech sampled at 16Hz. To ensure the accuracy of transcription, a two-stage alignment process was implemented. In this study, the train-clean-100 dataset is being selected, which contains a training set of 100 hours of "clean" speech [9].

Although the Common Voice dataset is large and open, there are significant differences in its characteristics compared to more curated corpora like LibriSpeech. Common Voice is a crowd-sourced dataset, resulting in a wide range of recording environments, microphone qualities, and speaker diversity. In contrast, LibriSpeech provides professionally segmented, high-fidelity, controlled speech. LibriSpeech serves as a stable benchmark to assess model performance under clean conditions, and Common Voice offers a realistic testbed to evaluate the robustness and adaptability of the model.

2.2. Model architecture

In this study, Wav2Vec2.0 employed a cutting-edge self-supervised approach for speech representation learning introduced by Alexei Baevski and other team members in 2020[5]. It gives a framework of how to pretrain models on massive amounts of unlabeled speech audio followed by

DOI: 10.54254/2755-2721/158/2025.TJ23214

fine-tuning these models on much smaller labeled datasets. The WavVec2.0 architecture relates to the extraction of rich contextualized representations directly from raw audio waveforms through a convolutional feature extraction/Transformer-based sequence modeling approach.

The overall model comprises three main parts: a feature encoder, a context network, and a quantization module. The feature encoder is a seven-layer convolution network that encodes raw audio input into latent representations at a lower temporal resolution. Those representations are then randomly masked and fed into the 12-layer transformer-based context network.

For the quantization module, one of the main difficulties of using Transformers for speech processing is the continuous nature of speech. Speech doesn't have natural sub-units. The quantization module uses a set of codebooks, each containing a fixed number of discrete vectors. These codebooks are learned during training. Wav2Vec2.0 is trained using a contrastive loss that encourages the model to differentiate the true quantized latent representation from a sample of distractors. From the original paper: "Wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned" [5].

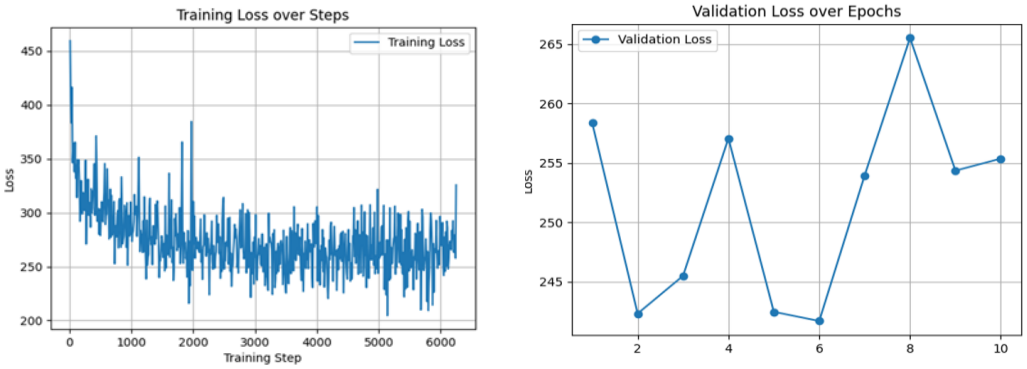
In the fine-tuning stage, a linear projection layer is added at the top of the pre-trained model, which maps contextualized embeddings to a vocabulary of characters or phonemes. Finally, the model is trained end-to-end using Connectionist Temporal Classification (CTC) loss.

For the experiments conducted in this work, the model selected was the facebook/wav2vec2-base-960h model in the Hugging Face Model Hub [10]. It contains approximately 95M parameters and is pre-trained on 960 hours of English read speech from the LibriSpeech corpus. The base model has 12 Transformer encoder layers with a hidden size of 768 and 8 attention heads. The pre-trained model is fine-tuned using domain-specific information from Mozilla Common Voice and LibriSpeech.

2.3. Results

Throughout the training, the training loss, validation loss, and validation WER/CER are obtained.

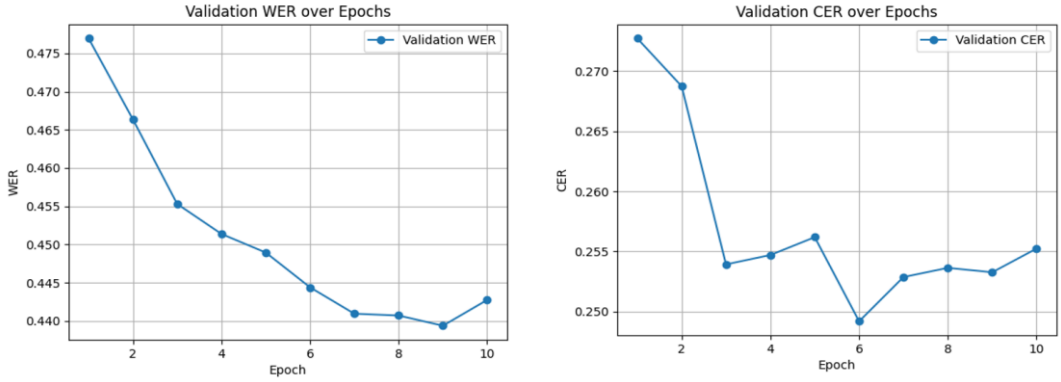
2.3.1. Mozilla Common Voice



(a) Training Loss Over Steps

(b) Validation Loss over Epochs

DOI: 10.54254/2755-2721/158/2025.TJ23214

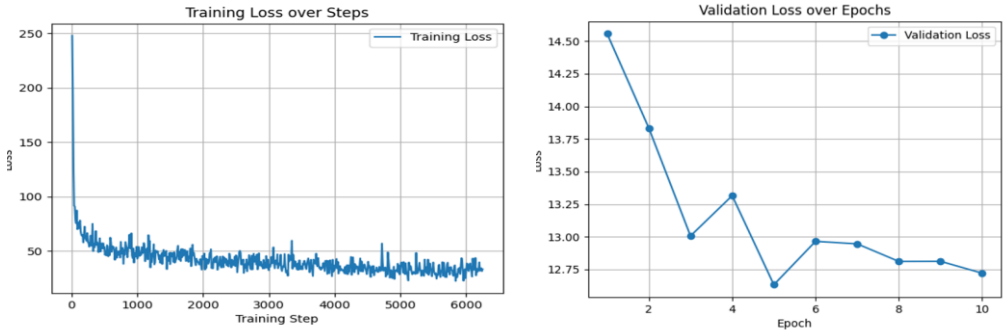


(c) Validation WER over Epochs (d) Validation CER over Epochs

Figure 1: Wav2Vec2.0 training performance on Mozilla Common Voice (photo/picture credit: original)

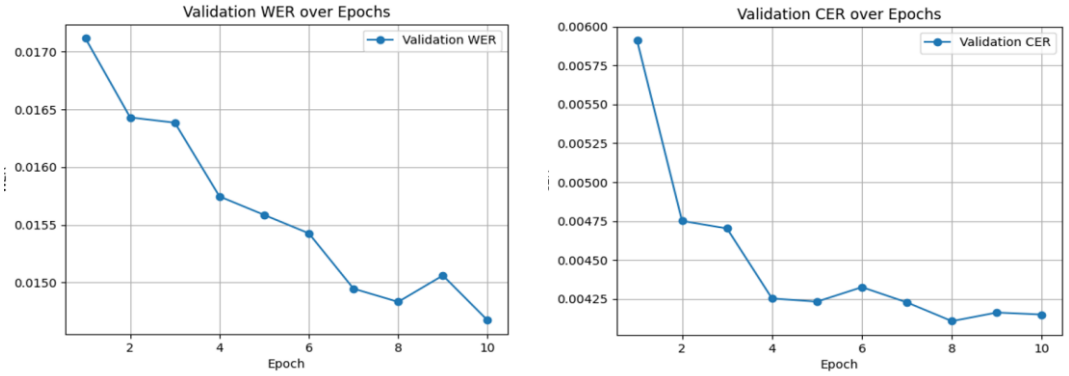
The Wav2Vec2. 0 is trained on the Mozilla Common Voice dataset for 10 epochs (Figure 1). Both the training and evaluation curves show that the training loss decreased monotonically and indicates that convergence during the optimization process was consistent. Validation loss (a common metric for evaluating model performance) varied among epochs. On the other hand, Word Error Rate (WER) and Character Error Rate (CER) have witnessed noticeable improvement. The WER on the validation set dropped from 0.476 at epoch 1 to 0.440 at epoch 10, with the final test set reaching a WER of 0.45. Another example is the decline in CER from approximately 0.272 to 0.255 through the training process. This means that despite the variability and noise in the Common Voice dataset, the model was able to generalize and maintain a steady performance. The fluctuations in loss metrics reflect the challenges of real-world speech data, but the consistent downward trend in WER and CER highlights the model's robustness and adaptability to diverse environmental conditions.

2.3.2. LibriSpeech



(a)Training Loss Over Steps (b)Validation Loss over Epochs

DOI: 10.54254/2755-2721/158/2025.TJ23214



(c) Validation WER over Epochs (d) Validation CER over Epochs

Figure 2: Wav2Vec2.0 training performance on LibriSpeech (photo/picture credit: original)

The wav2vec 2.0 model, fine-tuned on the LibriSpeech dataset, delivered a strong performance according to all evaluation metrics. Based on the resulting Figure 2, the training loss decreased over time, which is a sign of stable and effective learning during the training process. The validation loss decreased continuously throughout the epochs, indicating better generalization on training. In particular, the validation Word Error Rate (WER) decreased from about 0.0171 at epoch 1 to approximately 0.0147 at epoch 10. The Character Error Rate (CER) also decreased significantly, plateauing after a few epochs below 0.0043. The final test WER was 0.0153, which clearly shows the accuracy of the model on clean, well-segmented speech data.

2.4. Improvement

In addition to the result, there are directions for further improvement for both the training process and the model itself. First, incorporating domain-specific data augmentation strategies will be better to simulate the environmental variability, especially for crowd-sourced datasets like Common Voice. Additionally, in this research, the dataset is only about 1000 hours in total; using a larger dataset should enhance the model's capability and minimize performance degradation from noise data. From the model side, leveraging more advanced decoding techniques, such as language model integration or complete word dictionaries, should reduce the recognition error. Based on the training result on Common Voice, the CER is significantly higher than WER, which indicates that incorporating word decoding should help increase the WER metric. Overall, Wav2Vec2.0 shows great promise for scalable, accurate ASR, and with further optimization, it can be pushed closer to human-level transcription performance in diverse real-world applications.

3. Conclusion

In conclusion, this research demonstrated the effectiveness of the Wav2Vec2.0 model in ASR tasks using both the Mozilla Common Voice and LibriSpeech datasets. The model achieved a final Word Error Rate (WER) of 0.45 on Common Voice and 0.0153 on LibriSpeech, respectively. This result illustrates that the Wav2Vec2.0 model is robust and can adapt to diverse acoustic conditions. Nevertheless, the performance gap shown between Mozilla Common Voice and LibriSpeech is still huge, which should be caused by the different characteristics of the two datasets. LibriSpeech has clean, well-segmented recordings that provide an ideal setting for model fine-tuning, while Common Voice poses a more challenging, real-world environment, as the audio is collected through crowd work, in which speaker diversity, low recording quality, and unrelated noise might exist.

DOI: 10.54254/2755-2721/158/2025.TJ23214

References

[1] *Lu, X., Li, S., and Fujimoto, M. (2020). Automatic Speech Recognition. In Y. Kidawara et al. (Eds.), Speech-to-Speech Translation. SpringerBriefs in Computer Science. Springer, Singapore.*

[2] *Jain, A.D. and Ali, S.M. (2021) A Comparative Study of Speech Recognition Systems. 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE, pp. 303–307.*

[3] *Baevski, A., Schneider, S., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. arXiv preprint arXiv:1904.11660.*

[4] *Moreau, P. and Yvon, F. (2016) Evaluating Automatic Speech Recognition Systems in Comparison to Human Perception. HAL Archives-ouvertes, hal-01350057. https://hal.science/hal-01350057/*

[5] *Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv preprint arXiv: 2006.11477. https://arxiv.org/abs/2006.11477*

[6] *Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F.M. and Weber, G. (2020) Common Voice: A Massively-Multilingual Speech Corpus. arXiv preprint arXiv:1912.06670. https://arxiv.org/abs/1912.06670*

[7] *Common Voice Datasets. (n.d.). Retrieved from https://commonvoice.mozilla.org/en/datasets*

[8] *Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015) Librispeech: An ASR Corpus Based on Public Domain Audio Books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964*

[9] *LibriSpeech ASR corpus. Retrieved from https://www.openslr.org/12*

[10] *wav2vec2-base-960h. Hugging Face. Retrieved from https://huggingface.co/facebook/wav2vec2-base-960h*