Unveiling the Power of Data Augmentation: Enhancing ASR Performance with the Fine-Tuned Whisper Model

Po-Han Lu | Bachelor
- Department of Cross College Elite Program
- F14081088

Shao-Man Lee | Assistant Professor
- Miin Wu School of Computing, National Cheng Kung University
- Advisor

Abstract

This research examines the impact of data augmentation [1] techniques on ASR performance using the Whisper model [2]. Specifically, TimeStretch and PitchShift techniques are explored for their ability to reduce Word Error Rate (WER) [3] and Connectionist Temporal Classification (CTC) [4] loss. Specifically, TimeStretch and PitchShift techniques are examined, while the Gain technique is found to be ineffective and potentially detrimental to ASR accuracy. The findings highlight the importance of selecting appropriate augmentation methods that simulate natural speech variations and avoid artificial artifacts that can degrade ASR performance. This study emphasizes the importance of data augmentation in optimizing ASR tasks, contributing valuable insights for future advancements in speech recognition technology.

Related Work

Automatic Speech Recognition

Automatic Speech Recognition (ASR), is a technology that converts spoken language into written text. ASR systems analyze audio signals to extract meaningful information and transcribe it into textual form. The primary objective of ASR is to enable computers and devices to understand and process human speech, facilitating voice-controlled interactions, transcription services, voice assistants, and other applications. Recent advancements in deep learning have significantly enhanced the accuracy and usability of ASR.

Evaluation Metrics

CTC loss

CTC, Connectionist Temporal Classification, is a method employed in ASR. It offers a means to train models without requiring a direct alignment between the input speech and the resulting transcriptions. CTC introduces the "blank" symbol, which allows for the inclusion of repetitions and insertions in the transcriptions. The CTC loss function evaluates the probability of the correct transcription by considering the predicted probabilities and the expanded transcription. By minimizing the CTC loss, the ASR model gradually learns to align the acoustic units with the accurate sequence of words or characters in the transcription.

WER

Word Error Rate (WER) is an evaluation metric in ASR systems. WER calculates the percentage of errors between recognized words and the ground truth words. It provides a quantifiable measure of an ASR system's accuracy, where a lower WER indicates higher precision.

WER = (S+D+I) / N

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference (N=S+D+C)

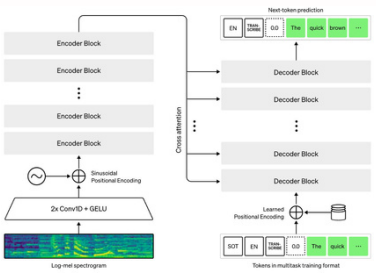
Experiment

Whisper introduction

The Whisper model is an automatic speech recognition (ASR) system developed by OpenAI, which has been trained on a diverse dataset comprising 680,000 hours of multilingual and multitask supervised data. This vast training dataset has led to improved robustness in the face of accents, background noise, and technical language.

Whisper Architecture

The whisper model utilizes an encoder-decoder Transformer architecture. Whisper divides input audio into 30-second segments and converts them into log-Mel spectrograms, which are then fed into an encoder. Decoder is trained to generate relevant text captions for each audio segment. By incorporating specific tokens, the decoder is able to handle different tasks, such as language identification, multilingual speech transcription, and speech translation into English. Through integrating these specialized tokens, Whisper becomes proficient at performing a wide array of tasks and delivering precise and adaptable results.



Motivation

This research aims to optimize the performance of the Whisper model in ASR tasks, addressing challenges posed by speech variations and diverse acoustic conditions. This research focuses on two key aspects: simulating variations through data augmentation during training and optimizing model parameters via fine-tuning to identify effective strategies for reducing Word Error Rate (WER) and loss. This research enhances the model's robustness and generalization capabilities, which contribute to advancing ASR technology and provide insights for practical optimization of the Whisper model.

Experiment

Data Augmentation for Audio Data

Gain

The "gain" augmentation involves adjusting the volume or amplitude of the sound. By increasing or decreasing the gain, the audio data can simulate different sound intensities.

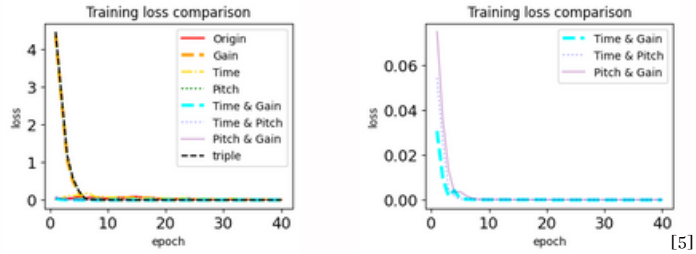
TimeStretch

"TimeStretch" augmentation alters the temporal properties of the audio. It allows stretching or compressing the duration of the sound without affecting the pitch, which can simulate various speaking rates or tempo variations in the dataset.

PitchShift

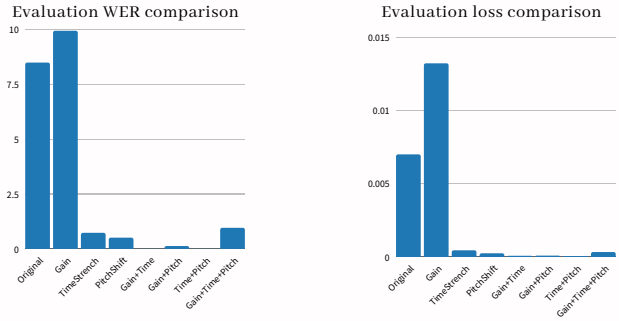
The "PitchShift" augmentation modifies the pitch or frequency of the audio while maintaining the duration. It can simulate changes in the pitch of the speaker's voice.

Results: Training Loss Function



[5]

Results: Evaluation



Conclusion & Discussion

Adjusting audio gain or volume lacks meaningful speech-related or acoustic variations, failing to simulate real-world speech variations. Moreover, it may introduce distortions, excessive noise, which cause negative impact on the performance of the ASR model and degrading accuracy instead of enhancing it.

Real-world situations often involve speakers with diverse speaking rates. TimeStretch allows to either stretch or compress the audio, making it longer or shorter while maintaining the same speech content, which enables the introduction of speaking rate variations, enabling the model to acquire knowledge and adjust to different speech speeds encountered in real-world scenarios.

Speakers exhibit diverse pitch ranges and vocal characteristics. PitchShift augmentation alters frequency, preserving speech content while simulating pitch variations. This enhances model robustness to voice pitch changes, ensuring accurate transcription regardless of vocal characteristics or intonations.

Reference

[1] A Survey of Data Augmentation for Audio Classification. Lucas Ferreira-Paiva, Elizabeth Alfaro-Espinoza, Vinicius M. Almeida. 2023
[2] Introducing Whisper, from https://openai.com/research/whisper
[3]An information theoretic measure of sequence recognition performance. Andrew C. Morris., 2006.
[4] Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J., 2006.
[5] SR-Generated Text for Language Model Pre-training Applied to Speech Tasks. Valentin Pelloin, Franck Dary, Nicolas Hervé. 2022