Report

1. Literature

Automatic Speech Recognition (ASR) technology has existed for a while now. It is helpful in many applications like in voice assistants, biometric identification, home automation by custom commands, video captionin, etc. One major application is also in the medical domain. For example, transcribing conversations between doctors and patients, or in capturing patient notes.

ASR is a technology by which computers can transcribe human speech to text. Traditionally, statistical models like HMMs were used for ASR. Now, with advancements in neural networks and deep learning, advanced methods like CNNs, RNNs, LSTMs and Transformer architectures are being used. These models have led to significant improvement in ASR accuracy.
Clinical ASR has certain challenges.

- Different vocabulary
- Scarcity of large datasets to train deep learning models
- background noise
- speaker variability

ASR models typically need a lexicon and have an acoustic model and a language model. Acoustic models learn the speech signals by learning features like MFCCs and extracting phonemes. Language models bridge the gap between the speech information and their meaning. For supervised training, the dataset should contain paired speech and transcripts. Preprocessing of transcripts to keep the relevant information, which would be beneficial for language modelling, is essential. End-to-end models exist, too. Different deep models exist for learning speech and audio representations, and these representations can also be considered as acoustic models. A transformer decoder could be used as the LM.

Fine-tuning deep neural networks (DNNs) for domain-specific data would help to improve the accuracy because DNNs are more robust to environmental and speaker variations and fine-tuning because data is less.

Dataset

PriMock57 is a conversational dataset between a doctor and a patient. It is a publically available dataset comprising of 57 mocked primary care consultations. Audio recordings

and their manual utterance-level transcriptions are available. In total, there are 7109 audio utterances and 6712 transcripts available. The audio samples without transcripts are ignored. The data is split into train, val and test sets in a ratio (80,10,10).

Dataset Preprocessing

The data is split into utterances, and utterance level transcripts along with file names are saved. Text preprocessing is also performed. Depending on the ASR method being explored, mel-spectrograms are extracted if needed.

2. Approaches studied

For zero-shot evaluation on the dataset, I utilised the following methods.

Wav2Vec ASR : Developed by Facebook AI, utilises self-supervised learning to map audio signals to an acoustic representation, and then it is finetuned with CTC loss to perform the ASR task. It is loaded using the hugging-face model (https://huggingface.co/osanseviero/asr-with-transformers-wav2vec2). Fine-tuning script is also implemented.

Whisper ASR: Developed by openAI, trained on a very large corpus for ASR tasks. It is a transformer encoder-decoder architecture with sequence-to-sequence learning objective. Loaded by referring to the official GitHub implementation (https://github.com/openai/whisper).

NeMo Conformer: Developed by Nvidia, based on Fast Conformer-based models. It is an end-to-end model. It is loaded using Nvidia's NeMo toolkit (https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/asr/intro.html).

3. Hypothesis

Since the data is conversational, large ASR models may not perform well if we consider the fillers while evaluating using word error rate (WER). But by fine-tuning the models, the performance may improve.

Since the amount of data available for fine-tuning is less, smaller models may perform better on the particular val-set.

4. Finding and results

Word error rates (WER) are calculated using a library called ‘jiwer’. WER is calculated in two cases, one considering the fillers (like “uh” or “Mmm”) and the other after removing the fillers. It is observed that pre-trained Whisper ASR models give best WER when fillers are not considered. Since fillers are not significant in the context of medical record keeping, it suffices if the ASR model is able to give good accuracy while ignoring the fillers. The following table shows the WER of the three methods on the test set. When I consider WER without fillers, I ignore the utterances that only have fillers in them. So, the number of samples for calculating WER in that case could be lesser.

Model	WER (no fillers)	WER (with fillers)	
Wav2Vec 2.0	0.666	0.429	Zero-shot
Whisper	**0.166**	0.343	Zero-shot
NeMo Conformer	0.5	**0.169**	Zero-shot