Research of Whisper by OpenAI

1. Understanding

Whisper is an automatic speech recognition (ASR) system trained by OpenAI with 680,000 hours of multilingual and multitask supervision. It is a speech-to-text technology that converts spoken words into written text. Training via large-scale weak supervision successfully balanced the trade-off between quality and quantity to achieve satisfying usefulness and robustness.

The sequence-to-sequence architecture of Whisper is a simple end-to-end approach, implemented as an encoder-decoder Transformer. The input sequence is fed into an encoder, processed, and compressed into a fix-size context vector, which is into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. The context vector from the encoder is then used as an initial state from the decoder. The decoder predicts the sequence of characters in the spoken text, one at a time. It uses the encoded representation from the encoder and its internal memory to autoregressively generate the output text. This architecture allows the model can perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.

As for the performance of the model, on the one hand, “the resulting models generalize well to standard benchmarks” (Radford et al., n.d.), which means it works well when recognizing languages like English. On the other hand, Whisper was trained on a large and diverse multilingual dataset, it performs well under diverse contexts without any specific fine-tuning or training on those particular datasets. According to Papers With Code, the Test WER of whisper-large is 5.4%, while the top model scores 3.06%. Multilingual LibriSpeech (MLS) dataset consists of 8 languages and doesn’t support Arabic (TensorFlow, 2022), while Whisper supports 57 languages (OpenAI, 2024). When OpenAI measures Whisper’s zero-shot performance across many diverse datasets it finds it is much more robust and makes 50% fewer errors than those models. The specifications of ASR API are below:

Criteria	Features
Languages Supported	57 languages
Rate Limit	50 requests /min
Audio Size	Up to 25MB
Pricing	$0.006 / minute (rounded to the nearest second)
Arabic WER/CER on Fleurs datasets	9.6%
English WER/CER on Fleurs datasets	4.1%

In conclusion, Whisper-large performs well in an English context. Moreover, it also suits a multilingual situation including Arabic to automatically generate the transcript between doctors and patients, which benefits our company's future international expansion.

2. Implementation

Setup:

```
# use pip to install Whisper
# compatible with Python 3.8-3.11 and recent PyTorch versions
pip install -U openai-whisper
# on MacOS using Homebrew (https://brew.sh/)
brew install ffmpeg
# may need to install rust as well
pip install setuptools-rust
```

Command-line usage

```
# model size: tiny, base, small, medium, large
whisper audio.flac audio.mp3 audio.wav --model medium
```

Python usage

```
import whisper

model = whisper.load_model("base")
result = model.transcribe("audio.mp3")
print(result[“text”])
```

References

OpenAI. (2024). *Whisper API FAQ | OpenAI Help Center*. [online] Available at: https://help.openai.com/en/articles/7031512-whisper-api-faq [Accessed 21 Jan. 2024].

Radford, A., Kim, J., Xu, T., Brockman, G., Mcleavey, C. and Sutskever, I. (n.d.). *Robust Speech Recognition via Large-Scale Weak Supervision*. [online] Available at: https://cdn.openai.com/papers/whisper.pdf [Accessed 22 Jan. 2024].

TensorFlow. (2022). *multilingual_librispeech*. [online] Available at: https://www.tensorflow.org/datasets/community_catalog/huggingface/multilingual_librispeech#:~:text=Multilingual%20LibriSpeech%20(MLS)%20dataset%20is [Accessed 21 Jan. 2024].