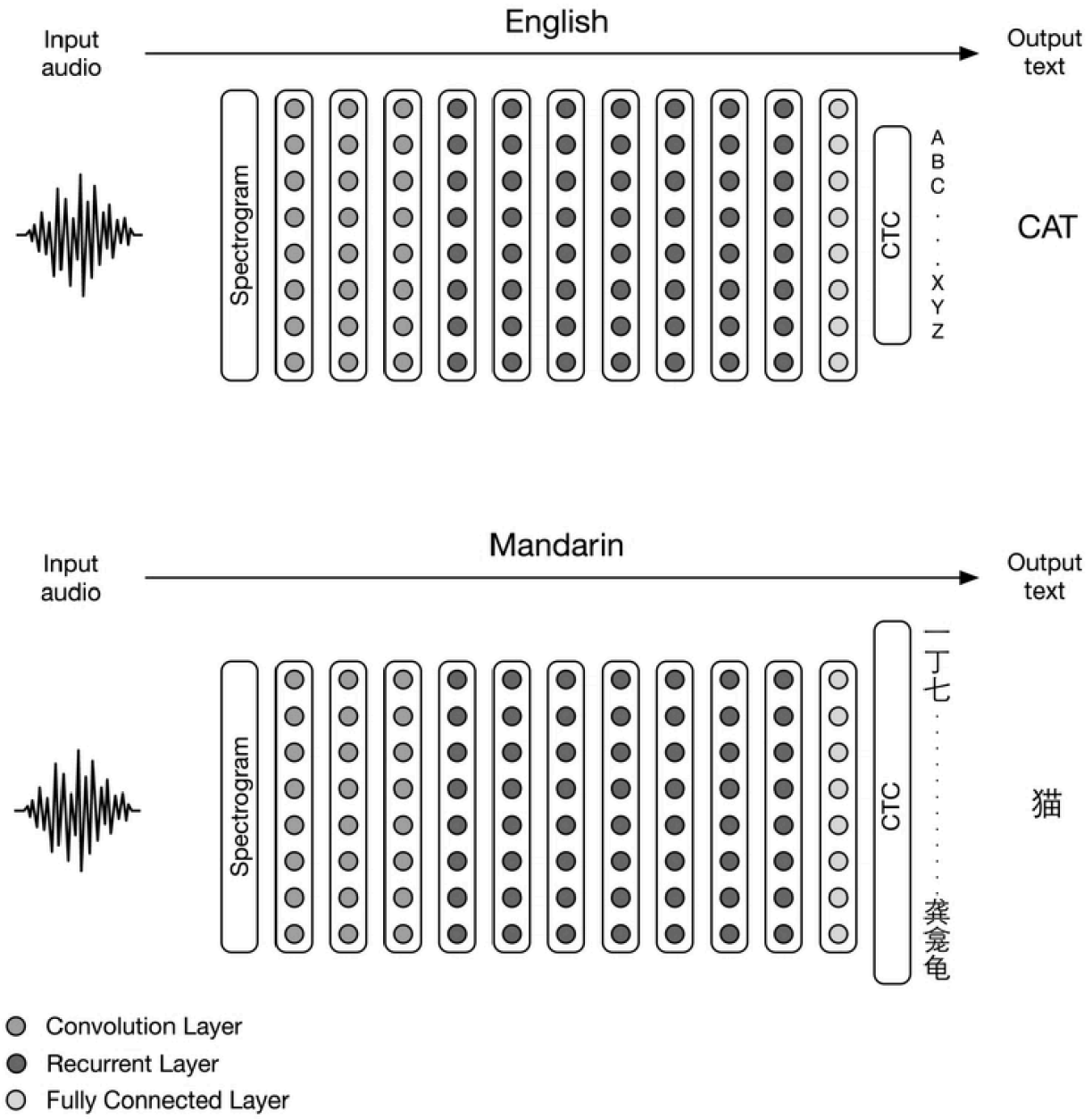
DeepSpeech

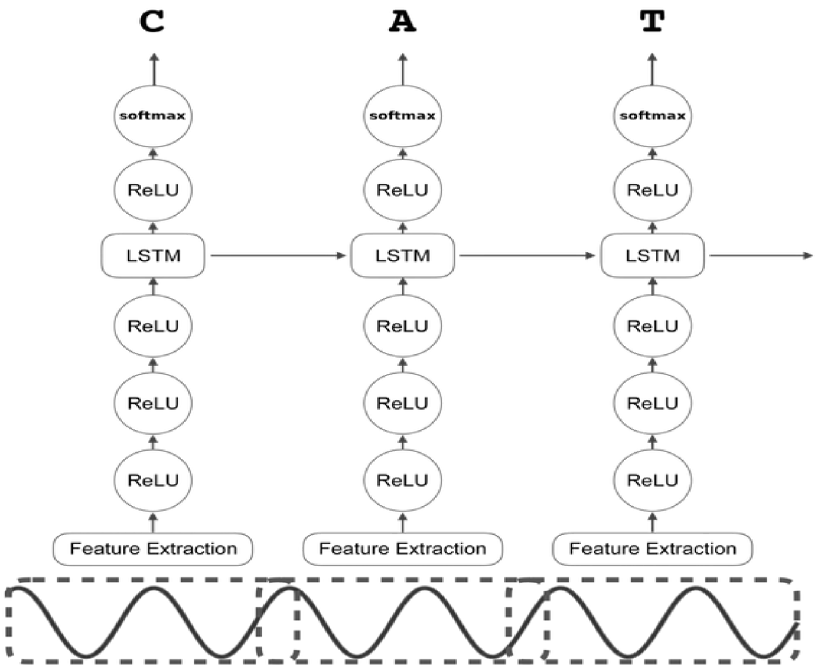
DeepSpeech is an open-source Speech-To-Text (STT) engine that uses a model trained by machine learning techniques. It was initially developed based on Baidu's Deep Speech research paper and is now maintained by Mozilla.



The engine is designed to automatically transcribe spoken audio into text. It takes digital audio as input and returns a "most likely" text transcript of that audio. This process is known as speech recognition inference. DeepSpeech can be used for two key activities related to speech recognition: training and inference. DeepSpeech is an implementation of an "end-to-end" speech recognition system. This means that the model takes in audio and directly outputs characters or words, as opposed to traditional speech recognition models that might involve multiple stages. The engine works by first converting the audio into a sequence of probabilities over characters in the alphabet. Then, this sequence of probabilities is converted into a sequence of characters. DeepSpeech can be used with appropriate hardware (like a GPU) to train a model using a set of voice data, known as a corpus. Then, inference or recognition can be performed using the trained model. DeepSpeech includes several pre-trained models. The engine is designed to run in real time on a variety of devices, ranging from a Raspberry Pi 4 to high power GPU servers. It can be installed and used via Python, and pre-trained English model files can be downloaded from the DeepSpeech GitHub repository. It's important to note that as of the latest updates, only 16 kilohertz (kHz) .wav files are supported.

How does DeepSpeech work?

DeepSpeech is an open-source speech recognition engine developed by Mozilla, based on the DeepSpeech algorithm by Baidu. It's designed to convert spoken audio into written text, a process known as speech recognition inference.



The core of DeepSpeech is a Recurrent Neural Network (RNN) that ingests speech spectrograms. The engine takes a stream of audio as input and converts it into a sequence of characters in the designated alphabet. This conversion involves two main steps:

1. The audio is converted into a sequence of probabilities over characters in the alphabet. This is achieved using a Deep Neural Network.

2. The sequence of probabilities is then converted into a sequence of characters.

The architecture of DeepSpeech includes several layers.

DeepSpeech can be trained using a set of voice data, known as a corpus. The training process involves updating the gradient to find the lowest loss, with the amount of processing done in one step depending on the batch size. By default, DeepSpeech processes one audio file in each step.

The performance of the model is evaluated using metrics like Word Error Rate (WER) and Character Error Rate (CER). WER measures how accurately DeepSpeech was able to recognize a word, indicating the performance of the language model (scorer). CER measures how accurately DeepSpeech was able to recognize a character, indicating the performance of the acoustic model.

References

1. https://klu.ai/glossary/deepspeech
2. https://deepspeech.readthedocs.io/en/r0.9/
3. https://arxiv.org/abs/1412.5567