Google vs. Facebook ASR and TTS model comparison

**Automatic speech recognition comparison:**

**Accuracy**: Google’s ASR service which runs on Google cloud has an accuracy of 95.9% according to their own website. The Facebook HuBert ASR model has been found to have an accuracy of 94% according to my own tests, and an accuracy of 97% according to Huggingface.co, an open-source community dedicated to making machine learning models public and easily interfaceable. While the Facebook model is good at recognizing words, it does not recognize context or and human sentences. The Google ASR can be used to add punctuation to sentences, and correct previous words based on current words by identifying context. For example, when saying “my son” and “the sun,” the Facebook model will return “my son” and “the son,” because it does not recognize context.

**Speed**: Google’s transcription of audio files processes about 2 seconds of audio data per second, while Facebook’s processes about .75 seconds of audio data per second. This makes the Google service over 2.6x father than the Facebook model.

**Flexibility**: Google’s model can serve just over 125 languages, while the Facebook model can serve only 16. Using Google’s service requires an internet connection because it utilizes google cloud, while Facebook’s model does not require an internet connection, because it is on-premises software.

**Price**: Google: $.016/minute of audio input. Facebook: Free and open source

**Overall**: Google’s cloud service beats the Facebook model is almost every category, as it is faster, supports sentence punctuation and context correction, and supports over 100 more languages. On the other side, Facebook’s model does not rely on any cloud service, is free and open source, and can compete well with Google’s accuracy.

**Speech synthesis comparison:**

**Accuracy**: Google’s: Easy to understand, no inconsistencies. Correct pronunciation of words 98% of the time. Facebook: Has some inconsistencies with pausing for pronunciation and pronouncing numbers. The last syllable of a number can run together with the first syllable of the next word, so it can be hard to understand. The model does not pause for periods, also making sentences difficult to understand sometimes. Sounds very human-like. Accurate pronunciation of words 85% of the time.

**Speed**: Google’s speech synthesis takes only .035 seconds per word, while Fakebook’s speech synthesis takes a whopping .65 seconds per word. This makes the Facebook model 18 times slower than Google’s service. If this was being used for an application that required real-time speech synthesis, this would be a deal breaker.

**Flexibility**: Google’s model can serve just over 40 languages with over 220 unique voices, while the Facebook model can serve only 8 languages, with 8 unique voices. Using Google’s service requires an internet connection because it utilizes google cloud, while Facebook’s model does not require an internet connection, because it is on-premises software.

**Price**: Google: $4.00 USD per 1 million characters. Facebook: Free and open source.

**Overall**: As far as speech synthesis goes, Google completely outpaces the on-premises Facebook speech synthesis. Google’s is far more accurate, nearly 20 times father, and can support far more languages and voices than the Facebook model.