Lab 3: ASR and TTS

LT2216: Dialogue Systems
Erik Kolterjahn Kjellberg

March 4, 2025

A. Hard cases for speech recognition

The Australian band King Gizzard and the Lizard Wizard has its own little universe, sometimes referred to as the *Gizzverse*, with fictional characters, places and events appearing across albums. I decided to try the ASR system out on names of songs, albums, characters, and events in their discography, as well as on the band name itself. Some of these consist of combinations of English words otherwise not occurring together, such as "people vultures" and "nonagon infinity", others are taken from other fantasy universes, such as "the balrog", while some are (to my knowledge) completely new words, such as "Han-Tyumi". A lot of names were recognized but the names seen in table 1 were only occasionally or never picked up.

Name	Confidence score (avg.)
King Gizzard and the Lizard Wizard	0.41
Polygondwanaland	0.27
the balrog	0.43
people vultures	0.87
Han-Tyumi	0.13
flying microtonal banana	0.43
intrasport	0.13

Table 1: Some hard to recognize names and the average of three confidence scores of the ASR model upon attempting at predicting them.

Each name was attempted to be recognized three times and the average confidence score taken as a metric. Out of these seven names, the lowest confidence scores are reported for words which are either completely new or new compound words in the form of just one word (Intrasport, Polygondwanaland), while compounds of existing words with spaces in between each word (e.g. King Gizzard and the Lizard Wizard) get higher scores. Therefore, I think the biggest problem lies in the names that do not consist of already established English words, such as "Han-Tyumi". In order to improve the recognition for these words, the ASR model probably needs to be trained on new vocabulary. Currently, it will instead try to match a word like "Han-Tyumi" to a combination of some already known words, ending up with predictions such as "how tire you me".

VG part

After training the model on a text file that included each of these seven terms, the result was a model that was able to recognize some but not all of them. The terms "King Gizzard and the Lizard Wizard", "intrasport" and "flying microtonal banana" now got average confidence scores of 0.75 (an improvement of 0.34), 0.15 (an improvement of 0.02) and 0.69 (an improvement of 0.26) respectively, for the occasions when the model *predicted correctly*, which indicates that the model has improved its recognition of these terms, especially two of them. However, the model still made incorrect predictions for these names as well. For the other names, no improvement was seen, and the model almost never predicted correctly. This might be because it only trained on text and not voice data for this task, and without an explicit sound file, it might be difficult for the model to infer how these new terms are supposed to be pronounced and therefore recognized.

The following endpoint ID can be used to replicate the results:
`3cba899e-92a3-4ff5-b4bc-ff6c183199d3`