Dylan Massey

Dialogue Systems Report: Lab 3

Part A: Hard Cases for Speech Recognition (~ **half a page pure text without spacing and questions**)

Now, try similar cases and reflect on the outcome:

1. Can you think of any names of fictional places, people or objects that are not recognized? (Keep your final project in mind!)

Different Types of Foods are not recognized, especially when using their native pronunciation. One example is Kötbullar. Instead, one would need to use the term 'Swedish Meatballs'. As for names and places any fictional names that become reasonable complex (and are not compositions of terms familiar in the English language) appear to become difficult for the system to recognize.

One example: Just swapping the first two letters of my name 'Dylan Massey' to 'Mylan Dassey' will lead to erroneous transcription, i.e. Milambasi. Even a seemingly simple name changes such as 'Victoriu' instead 'Victoria' leads to problematic transcription. This hints to the fact that the transcriptions are probabilistic (corpora-based) rather than simply based on transcription of phonemes.

Cuss words appear to be transcribed but not returned in their outputs.

2. Are there any real location names or names of people that are also not properly transcribed?

One example is 'Zellig Harris', which is simply not recognized properly and often wrongly transcribed as 'Silly Caris', or 'Selikaris' or 'Celia Harris'. Similar problems occur for names such as 'Kazimierz Ajdukiewicz' or also 'Per Martin-Löf'.

For places the same applies, i.e. the German city of "Zwickau" is not properly transcribed and instead transcriptions such as 'TVKO' or 'Tzvikov' are suggested. On the other hand a place such as the German city of "Ravensburg" is more easily recognized. This makes sense to a degree, because both terms of the compound are actually recognizable terms in English and thus the pronunciation of the individual parts "raven" and "burg" appears to lead to more easy recognition.

3. Do you think any specific accent you are using makes words difficult to process?

I've noticed that 'dialing up' my American accent seems to improve recognition for some names. I.e., using an 'ultra fake American' pronunciation for 'Zellig Harris' leads to the almost correct transcription 'Zelig Harris'. When using a more British accent the transcription accuracy seems to fade.

4. Write some sample code to show confidence scores in speech recogniton, or you can log it in the console. How good are those scores?

I've decided to log them to the console. One noteworthy observation might be that the confidence scores are higher for correctly identified items vs. incorrectly identified items. There seems to be little "in-between". For example, the kind of fictional name "Mylan" (though Milan exists of course) is transcribed to "Meal in" with confidence of .059. When saying "Meal in" though the confidence spikes to .46.

5. Think about how this problem, transcription of something we did not intent to say, could be solved. Why do you think recognition falters in the examples that you tried?

If the transcription is unique and there is a low likelihood that the erroneous transcription will ever occur during interaction with the system, we could simply map it to the "actual" intended transcription. In this case we would kind of "patch" the terror made by the ASR-system. Alternatively, we could add some custom phonetic transcriptions of possible pronunciations of the words and allow the system to match them.

As to the reasons for the errors one might speculate that it has to do with the frequency of the words / subwords in the training corpora. As soon as the names seem reasonably unique, which arguably also is a question of how rare the names are within the English-speaking "world", the less likely it appears that the names are recognized. Another factor might be how simple it is for a given name to generate an adequate "English" pronunciation. 'Anna' for example as a German name is easily also pronounced for English speakers, other names are more difficult.

<u>Part A – VG:</u>

Which new words are now supported and can be tested. Report should contain your Endpoint ID (of the model).

I added the following terms to be recognized along with their UPS transcriptions:

- kazimierz/K A SH IH M IH R SH
- per/P EH
- mylan/M IH L AX N
- ajdukiewicz/AI D UH K EH V I D ZH
- zwickau/Z V I K AU
- dassey/D AE S I

More details can be found in the accompanying phon_prun.md found in ./Reports/lab3 of this repository. I also made use of some template sentences for better testing.

Astonishingly the model appears to have learned the custom names. Only Zwickau seems to be quite difficult for the system to recognize. With some concerted effort from my side though – namely voicing the 'z' very explicitly – the recognition appears to work.

The endpoint ID is: 0c05d96d-0c83-4f9f-bcd9-1a307ce63700

Part B

Poem choice: 'Do not go gentle into that good night.'
https://www.youtube.com/watch?v=w-sM-t1KI_Y&t=71s

Written version:
https://poets.org/poem/do-not-go-gentle-good-night

I chose the following couple of verses:

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

Though wise men at their end know dark is right,
Because their words had forked no lightning they
Do not go gentle into that good night.

Notes to Part B:

- No English Welsh accent is unfortunately available, but Michael Sheen was born in Wales. So I chose an alternation between Irish and British OpenAI voice, just to make it a little more different from the Standard Southern British English.
- There is a lack of control for OpenAI chosen voices when it comes to pronounciation control. A trick here is to first choose an non multilingual voice, then edit the phonetics and then switch back to the multilingual voice.
- Can we elongate specific vowels, which is used frequently in reciting poems. An example is controlling the vowel length of "burn" (/bɜːn/), where it would be desirable to add temporal control to the pronounciation time of **ɜ.** At least we can put the stress on the aforementioned phoneme.
- In general, I tried to add adequate pauses and elongate the vowels accordingly.