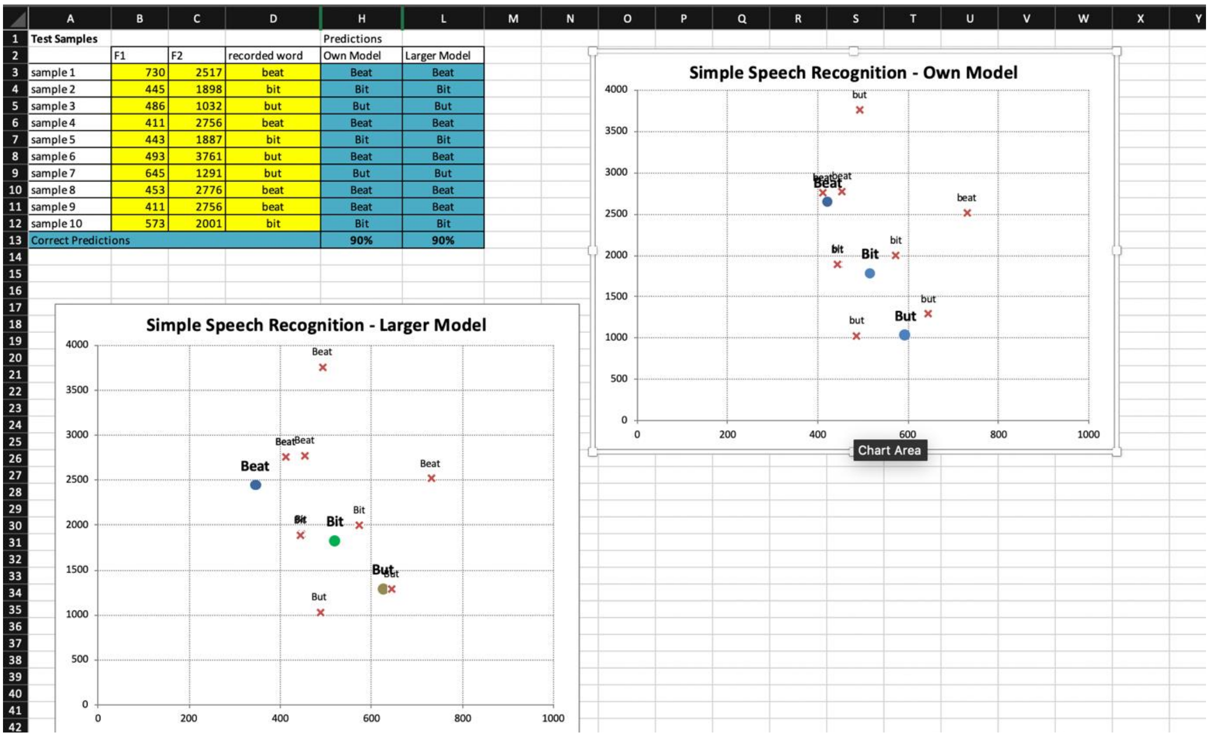
Part 1

a) Describe your experience and your results (use screenshots where necessary). Which model performed better? Which words were most often confused?



Test Samples				Predictions	
	F1	F2	recorded word	Own Model	Larger Model
sample 1	730	2517	beat	Beat	Beat
sample 2	445	1898	bit	Bit	Bit
sample 3	486	1032	but	But	But
sample 4	411	2756	beat	Beat	Beat
sample 5	443	1887	bit	Bit	Bit
sample 6	493	3761	but	Beat	Beat
sample 7	645	1291	but	But	But
sample 8	453	2776	beat	Beat	Beat
sample 9	411	2756	beat	Beat	Beat
sample 10	573	2001	bit	Bit	Bit
Correct Predictions				90%	90%

Both models preformed equally good. This might be due to the fact that I only used male voices without a thick accent so they sounded similar to me.
When I used a Chinese person ,which had a thick accent, as a sample both models failed to successfully identify the word (sample 6).

b) In your opinion, when and why are larger training sets necessary? What are the drawbacks/benefits of using recordings from a group of people as opposed to several examples from the same person? Can you think of ways the model could make better use of the larger amount of data rather than simply averaging over the samples?

Larger training set are necessary when we want to open the model to the general public. they are necessary because different people speak differently ,for instance females tend to have a voice that has a higher pitch.

The drawback of having a group of people instead of a single person is that there will be people that have voice that is too different to the others and will manipulate the data in the wrong direction. But the advantage of having a large enough group of people is that the model can predict more accurately for more people while the other model can predict accurately for the same person.

An alternative approach to this problem is to have several averages, e.g. one for females ,one for Irish people with a thick accent, one for Africans with a thick accents , one for Iranians and etc.

c) Would this method be useful to distinguish between the words beat and beet? How would you solve this problem?

No, because they sound exactly the same and this this model cannot understand the context around the word. The only way that I can think of to solve this problem is to have a big enough neural network that can understand and predict the meaning of a whole sentence.

Part 2

I used Pewdiepie's latest video am i a boomer as an example

a) a) What type of mistakes does the automatic transcription make? What kinds of things are transcribed correctly?

The caption is pretty much accurate until he invites a friend into the video at which point the friend talks in Swedish and the caption is stopped for a few seconds and it doesn't even write a subtitle when his friend starts talking English, it's worth mentioning neither of these people has an accent. When Pewds and his friend talked at the same time the caption stopped generating for one of them, but I had no trouble understanding both

It also transcribed G-fuel which is a beverage as 'gee fuel'

It also transcribed Christmas music as 'Crancky's music'

It transcribed "they are so entitled my gonna get things for free" instead of "they are so entitled why should they get things for free"

It transcribed "I should be illegal" instead of "that should be illegal"

b) Estimate what percentage of the content of the video was transcribed correctly.

I would assume more that 75 percent

c) When you examine the caption, do you find: - nonsense words, - spelling errors, - incorrect grammar, - phrases that don't make sense?

Yes, there are some instances of it, which is mentioned above, I would assume it's mainly because both of them talk at the same time. It may also be because of the fact that it cannot understand the context of the sentence

d) Which parts of the video appear to be more challenging for the speech recogniser and why?

The parts where they speak too loudly or laugh while the other person is speaking or when they both speak at the same time.

For "The Two Ronnies"' clip

a) What type of mistakes does the automatic transcription make? What kinds of things are transcribed correctly?

Similar to the shop owner the subtitle couldn't recognize the puns that were being made.

It transcribed 'fork' as 'folk'.

It also captioned "foot size" Instead of "What size".

b) Estimate what percentage of the content of the video was transcribed correctly.

More that 60 percent of this 2-minute clip was captioned wrong

c) When you examine the caption, do you find: - nonsense words, - spelling errors, - incorrect grammar, - phrases that don't make sense?

Yes, one instance of that is when the shop owner asks for the size of the rubber plug with question "what size" and the subtitle says "foot size"

Another is when the customer says “handles for forks” and it is transcribed as “handle for folks”.

e) Which parts of the video appear to be more challenging for the speech recogniser and why?

The part where the accent is too thick, or people are laughing in the background obviously enough it cannot recgnize the puns either.

Part 3

a. What difference did you notice when reading the two bold words in examples 1 and 2?

We read each letter of the word MTV (with phonetics əmtɪvɪ) but read NASA as a single word.

b. How do you think this problem can be solved?

People can read NASA as (en ai es ai) otherwise we must separately program our model to understand these abbreviations.

What cues can you provide the synthesiser with to help decide how to pronounce the numbers in examples 3 and 4?

We can tell the synthesiser to change the words nineteen-eighty-four and a thousand nine-hundred and eighty-four to 1984 with a few if-statements.

Part 4

a) What difficulty do you think a machine would have reading these sentences aloud?

Th machine would not be able to differentiate between 2 words that have the same spelling if the machine cannot understand the context of the word. Therefore, live the verb and live the adjective would be read the same way

b) Think of some cues a system can use when deciding how to read a word.

They can check the grammatical place of the word for instance ‘live’ in the first sentence is an adjective while ‘live’ in the second example is a verb.

They can also use the context of the words around the word to guess the pronunciation. For instance, in the fourth example we have the verb “feeling content” we can develop a model that read content this way whenever the word “feeling” or “feel” or “feels” is around it.

We can also use the contextualization method for the last example as well, we can manipulate the model to read tear as “tɪər” whenever we have the word drop around it.