Error Prediction on Speech Recognition using NLP

Abhishek Verma

Table of Content

Overview: Speech Recognition Errors

- Various speech to text conversion tools are available as opens source packages that can be implemented for professional and personal use.

- Their main function is to take input audio clips and convert to text in desired languages.

Overview: Speech Recognition Errors

- The quality of audio content, background noise level, language accent, speaker voice etc all plays a big role in the accuracy of the text content perceived.

- This means that some errors and uncertainties in the textual output are generated as a result of poor audio quality.

Objective: Error Prediction

- The purpose of this project is to construct a deep learning flow using NLP to predict the speech recognition errors generated as a result of bad audio quality .

- In particular the model will predict **wording error** on speech translated texts.

https://www.danielcraig.com/2010/05/09/dangers-of-error-correction-for-better-or-for-worse-comic-strip/

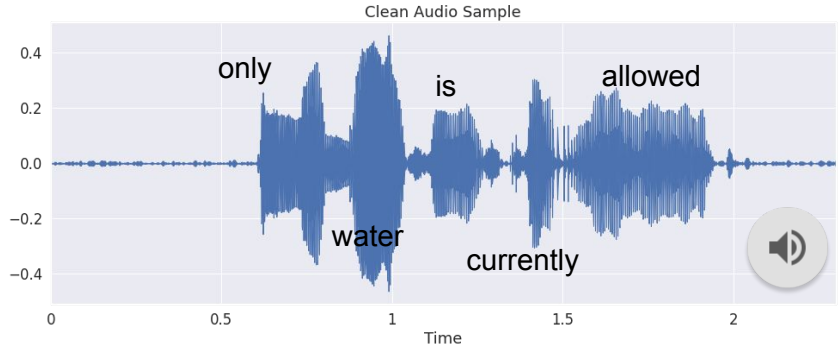
Data Source: Audio Clips

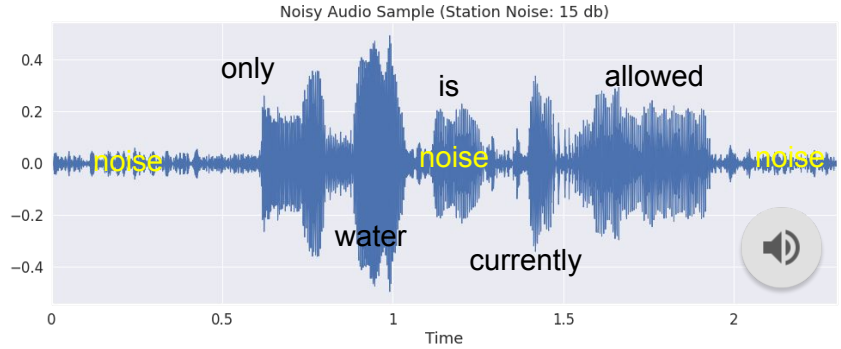
- I use the audio clips from clean and noisy parallel speech database used in the study "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks" by C. Valentini-Botinhao, X. Wang, S. Takaki & J. Yamagishi, In Proc. Interspeech 2016.

- References: https://datashare.is.ed.ac.uk/handle/10283/2791

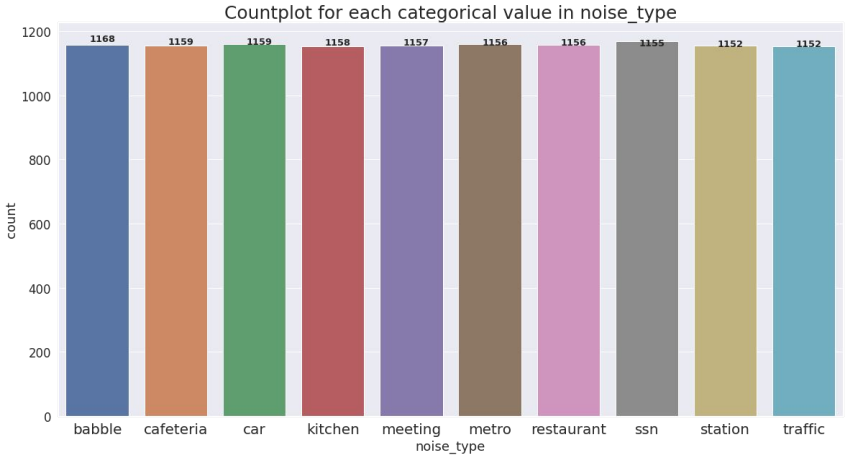
- These audio clip were downloaded from the website and accessed from local drive.

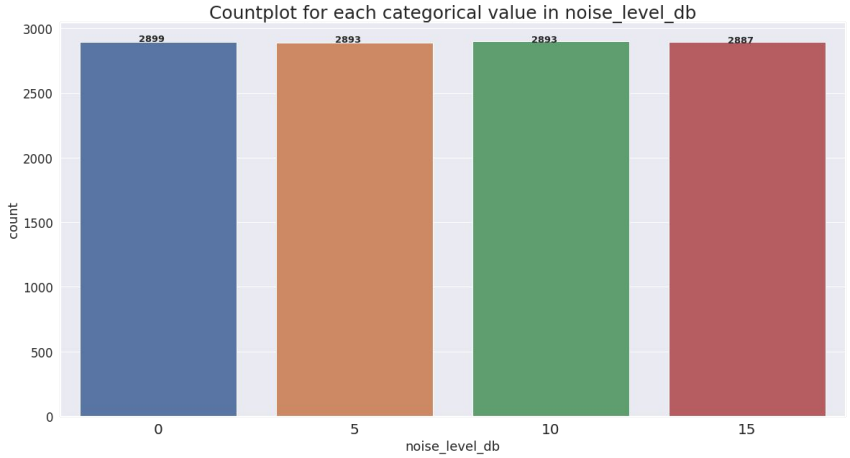
Data Source: Audio Clip Statistics





Total 2 x 11000 audio clips





Data Source: Speech Recognition Tool

- Python package to translate audio clips to text.

- For this project, I use the google web speech API.

SpeechRecognition 3.8.1

```
pip install SpeechRecognition
```

https://realpython.com/python-speech-recognition/

Real Python

Data Source: Clean Audio vs Noisy Audio

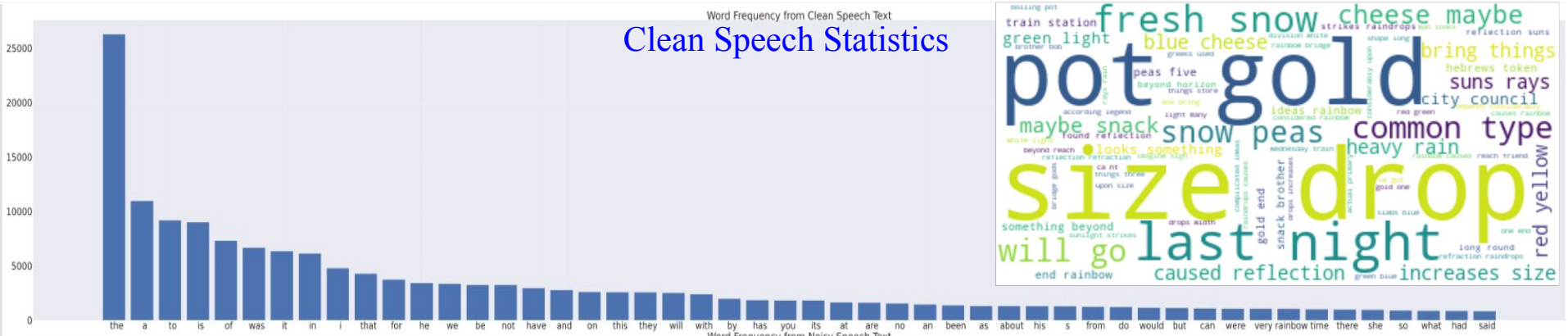
- Clean Audio Example

```
when the sunlight strikes raindrops in the absence of presidential in the rainbow
```

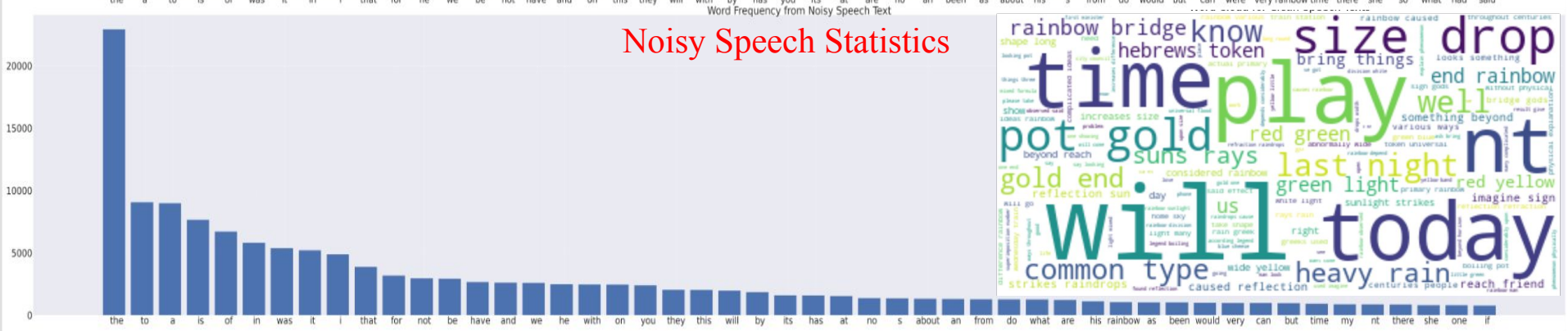
- Noisy Audio Example

```
striped raincoat
what is the Light Strike train coming
what is the Light Strike range
what is the light strike train timing
what is the Light Strike Crane Company
what is the light strike train cousin
what is the light strike train cannon
what is the light strike train kunming
what is the light strike ragecomic
```

Data Source : Word Frequency



Clean Speech Statistics



Noisy Speech Statistics

Data Transformation: Textual Cleaning

```
  Raw Text ----------> mr. Ferguson became a minister after 7 years is a generalist
Clean Text ----------> mr ferguson became a minister after seven years is a generalist
```

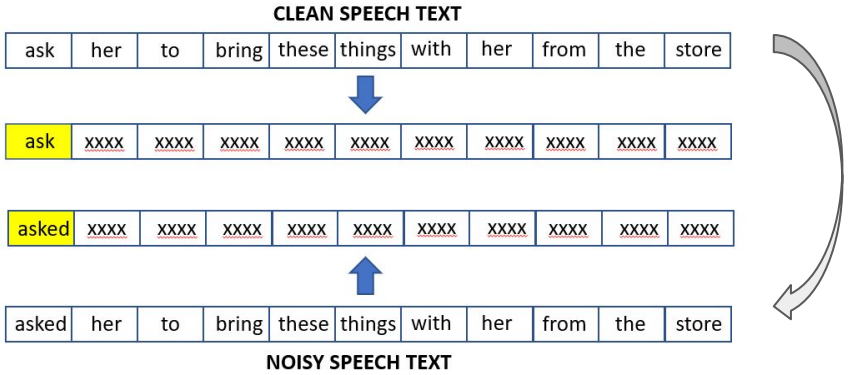
The following basic textual processing was applied to all the sentences:

- Omit punctuations in the words (comma, quote etc)
- Remove leading and trailing spaces between words
- Remove alpha-numeric texts
- Convert numeric "digits" to word representations
- Lower case all words

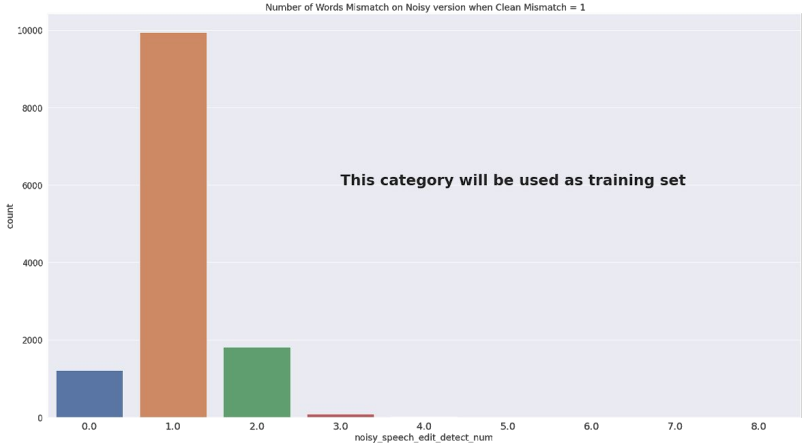
Data Transformation: Identify Error in Words

```
Cases where one word in clean was mis-translated to one word in noisy case

Clean Speech Version --> ask her to bring these things with her from the store
Noisy Speech Version --> asked her to bring these things with her from the store
Clean Speech Version with Detect --> ask xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx
Noisy Speech Version with Detect --> asked xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx
Clean Speech Version: No of word that dont match with noisy version --> 1.0
Noisy Speech Version: No of word that dont match with clean version --> 1.0

Cases where one word in clean was mis-translated to two words in noisy case

Clean Speech Version --> from that day on we started to look for another fullback
Noisy Speech Version --> from that day on we started to look for another full back
Clean Speech Version with Detect --> xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx fullback
Noisy Speech Version with Detect --> xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx full back
Clean Speech Version: No of word that dont match with noisy version --> 1.0
Noisy Speech Version: No of word that dont match with clean version --> 2.0

Cases where one word in clean was mis-translated to three words in noisy case

Clean Speech Version --> i did nt have a bath on myself
Noisy Speech Version --> i did nt have a bath on my cell phone
Clean Speech Version with Detect --> xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx myself
Noisy Speech Version with Detect --> xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx my cell phone
Clean Speech Version: No of word that dont match with noisy version --> 1.0
Noisy Speech Version: No of word that dont match with clean version --> 3.0

Cases where one word in clean was mis-translated to multiple words in noisy case

Clean Speech Version --> i feel i did not have enough time
Noisy Speech Version --> i did not have enough time
Clean Speech Version with Detect --> xxxxxx feel xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx
Noisy Speech Version with Detect --> xxxxxx did not have enough time
Clean Speech Version: No of word that dont match with noisy version --> 1.0
Noisy Speech Version: No of word that dont match with clean version --> 5.0
```

Mask all common words

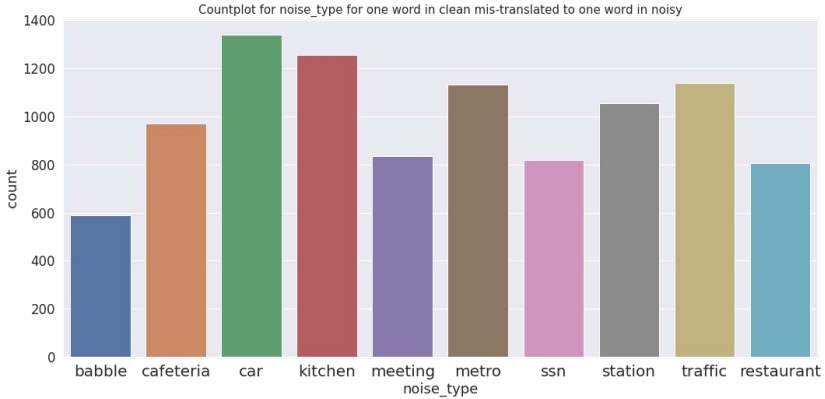


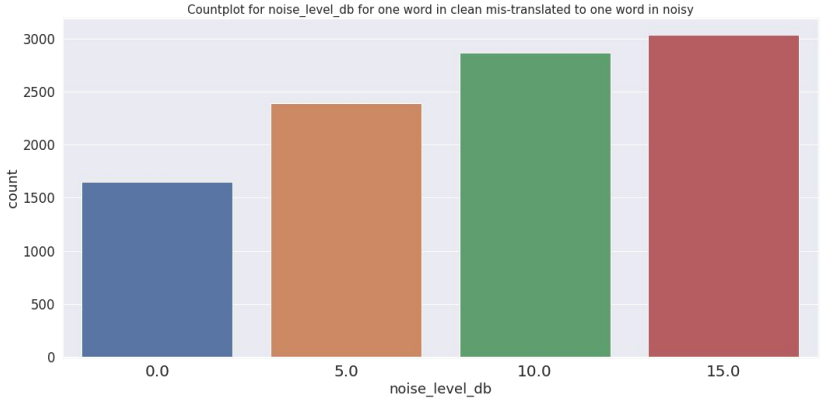
Feature Engineering: Data Selection



- Multiple words can be mistranslated as speech errors.

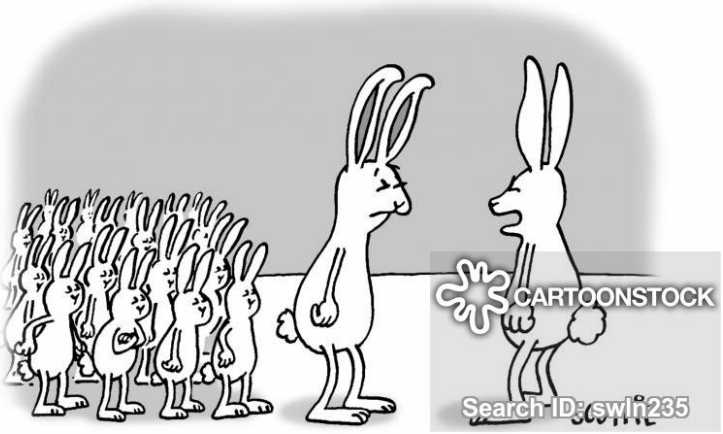
- In total 9939 sample with error in translating **one word**.





Feature Engineering: Sentence Subsampling

Word1 Word2 Word3 --- --- --- --- WordN



"There you go again, repeating yourself!"

#	Word1	Word2
Word1	Word2	Word3
Word2	---	---
---	---	---
---	---	---
---	---	---
---	---	---
---	WordN	#

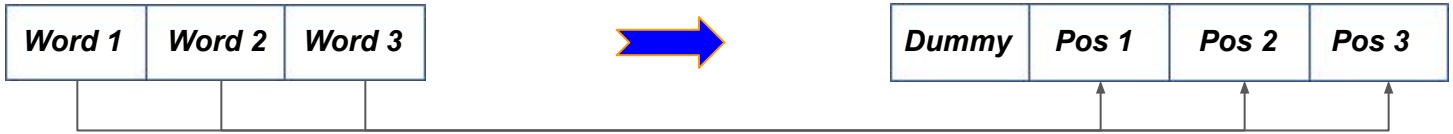
Feature Engineering: Input (Word Tokenizer)

Each unique word attached to a number

```
Input Word :     in the air      ---------->    [  8   2 708]  : Input Hotcoded Data
Input Word :     the air they    ---------->    [  2 708  24]  : Input Hotcoded Data
Input Word :     air they act    ---------->    [708  24 478]  : Input Hotcoded Data
Input Word :     they act as     ---------->    [ 24 478  33]  : Input Hotcoded Data
Input Word :     act as a        ---------->    [478  33   3]  : Input Hotcoded Data
Input Word :     as a prison     ---------->    [ 33   3 479]  : Input Hotcoded Data
Input Word :     a prison in     ---------->    [  3 479   8]  : Input Hotcoded Data
```

Feature Engineering: Output

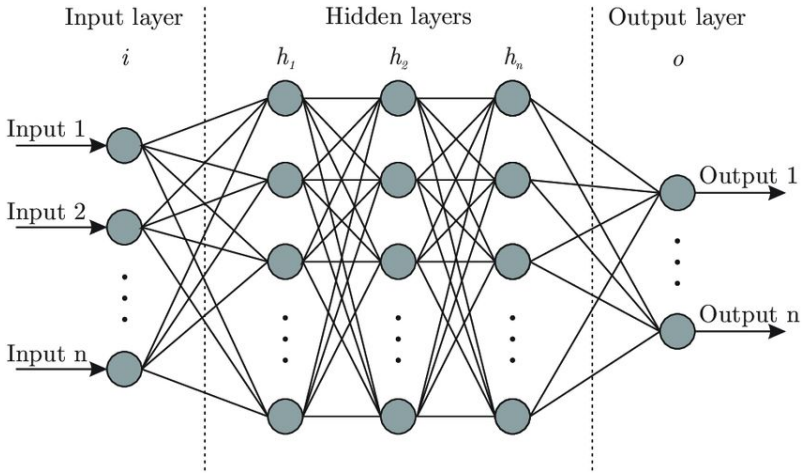
For no errors, 'Dummy' is activated




```
Input : # asked her ---------------------> [0, 0, 1, 0] : Output
Input : asked her to --------------------> [0, 1, 0, 0] : Output
Input : her to bring --------------------> [1, 0, 0, 0] : Output
Input : to bring these ------------------> [1, 0, 0, 0] : Output
Input : bring these things --------------> [1, 0, 0, 0] : Output
Input : these things with ---------------> [1, 0, 0, 0] : Output
Input : things with her -----------------> [1, 0, 0, 0] : Output
Input : with her from -------------------> [1, 0, 0, 0] : Output
Input : her from the --------------------> [1, 0, 0, 0] : Output
Input : from the store ------------------> [1, 0, 0, 0] : Output
Input : the store # ---------------------> [1, 0, 0, 0] : Output
```

Modelling: Deep Learning NLP Model

```
maybe the special is beating
what book did that take
today i could nt run on it
among them how many criminals
that is the face of boe
```

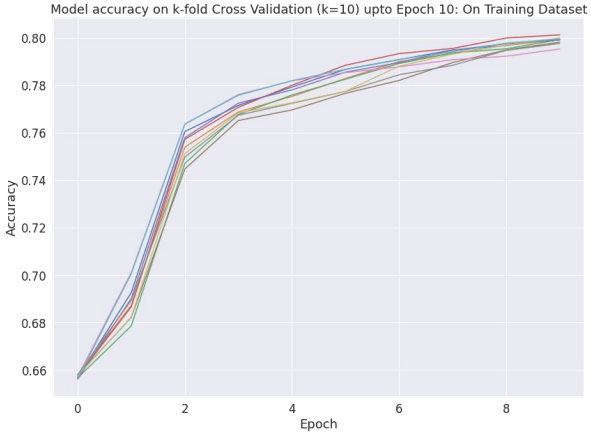


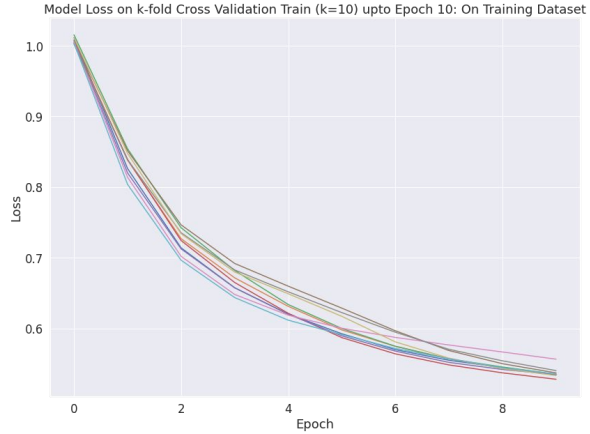

```
[{'beating': 0.67}]
[{'what': 0.67}]
['NO ERROR DETECTED']
[{'how': 1.0}]
[{'boe': 0.33}]
```

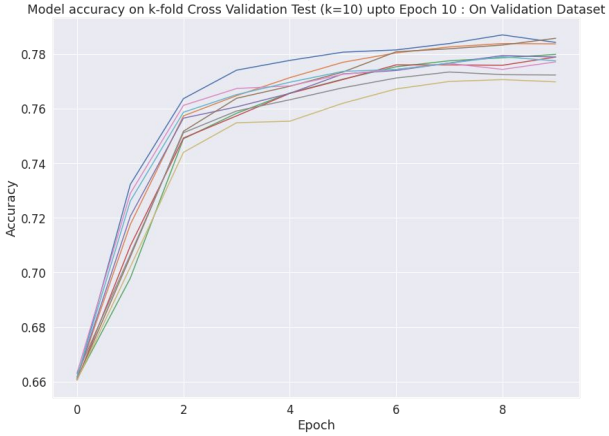
Modelling: K-Fold Performance

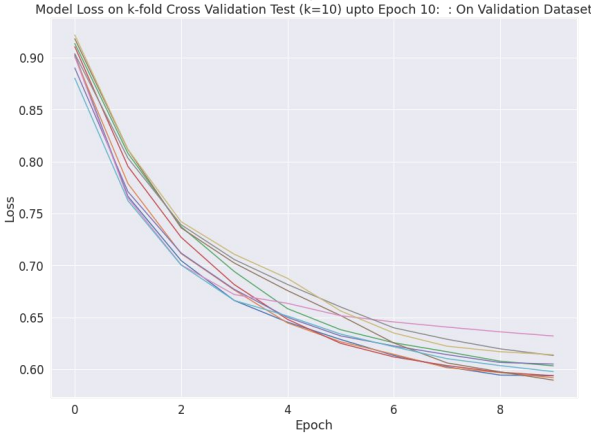
	validation accuracy	train accuracy
0	78.426254	81.173706
1	78.371328	81.278992
2	77.986819	81.021118
3	77.890688	81.584167
4	77.890688	81.201172
5	78.577316	81.004333
6	77.712166	80.586243
7	77.231532	80.847168
8	76.981187	81.338787
9	77.750307	81.038195

Mean Train 81.11% (+/- 0.26%)
Mean Validation 77.88% (+/- 0.48%)









Modelling: Improvements

- Data augmentation by adding rhyming word to modelling inputs.

Rhyming or Rapping

.....A	I'm chilling in this place
.....A	Look me in my face
.....A	I'm winning this race
.....A	You just a disgrace
.....B	I try not to boast
.....B	But I'm hotter than toast

	target	rhymcomb
0	[made, on, the]	made aune the
1	[made, on, the]	made on the
2	[the, scottish, parliament]	the british parliament
3	[the, scottish, parliament]	the skittish parliament
4	[the, scottish, parliament]	the scottish parliament
5	[the, scottish, parliament]	the fetish parliament
6	[the, scottish, parliament]	the brutish parliament
7	[the, scottish, parliament]	the smartish parliament
8	[the, scottish, parliament]	the lettish parliament
9	[the, scottish, parliament]	the smartish parliament
10	[the, scottish, parliament]	the fetish parliament
11	[the, scottish, parliament]	the skittish parliament

Conclusion

- Speech Recognition Errors can be successfully predicted using NLP and deep-learning modelling.

Questions