Error Prediction on Speech Recognition using NLP

Abhishek Verma

Table of Content

Presentation Duration : 20 min

Overview: Speech Recognition

- We are currently using many tools and instruments that operate on human voice based command. Most essential procedure to make this possible is to use the **speech to text conversion** routine.

- The main function here is to take input as audio signals and convert to textual information that can then be used in algorithms.

Overview: Speech Recognition having Errors ?

- However, in real world scenario, these voice commands are coupled with background noises. Sources of noise could be from the outside traffic, public chatters, road noise in cars etc.

- Can the noise interference have some effects on the textual interpretation? Do we have evidence to show?

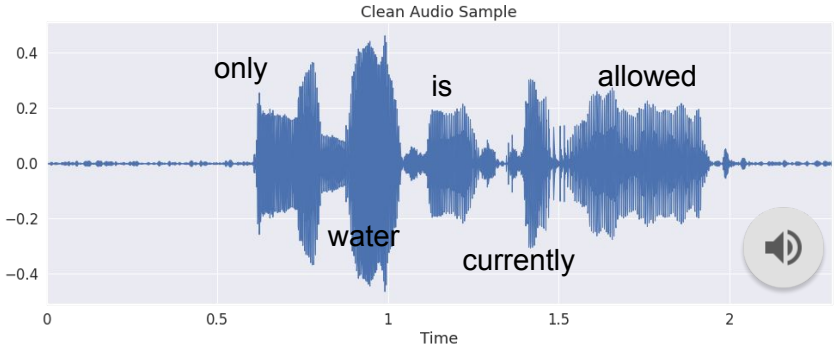
- If so, can we predict the errors using machine learning tools and apply correction to them? How accurate can we get ?

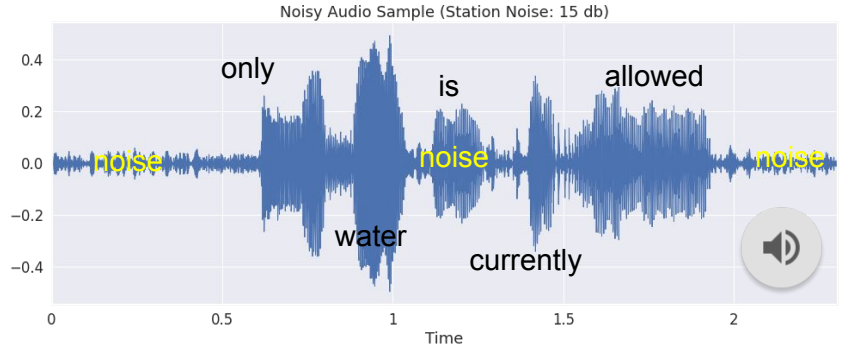
Data Source: Audio Clips

- 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.

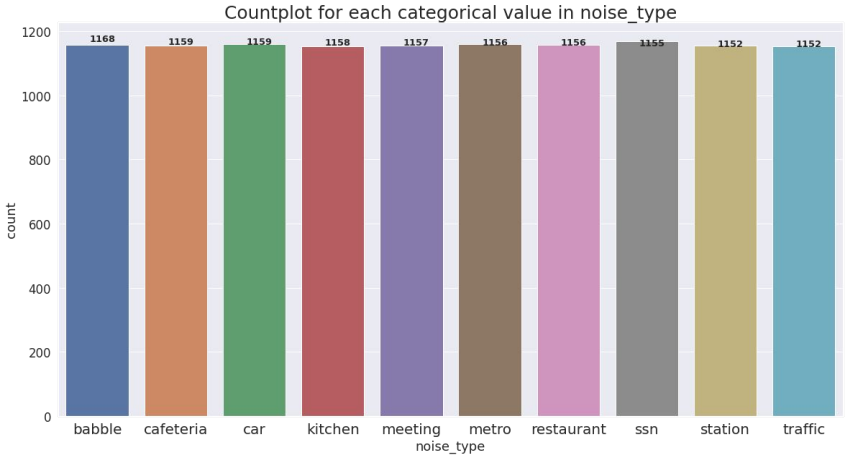
- References: https://datashare.is.ed.ac.uk/handle/10283/2791

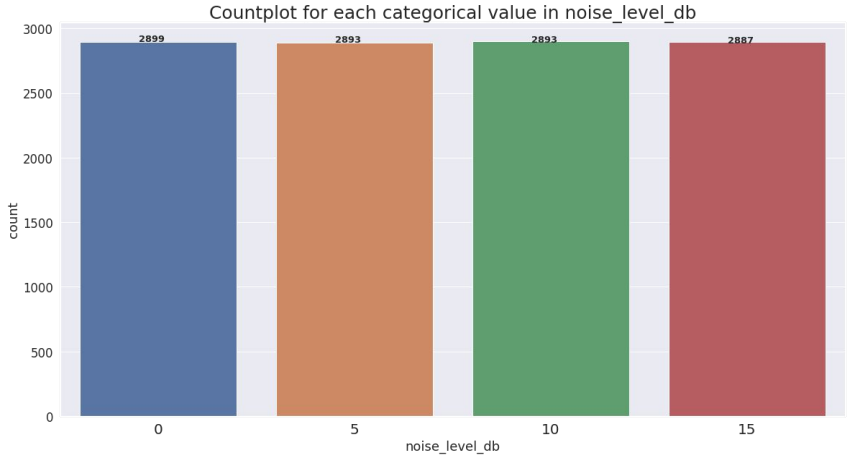
Data Source: Audio Clip Statistics





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
```

Speech Conversion: Clean Audio vs Noisy Audio

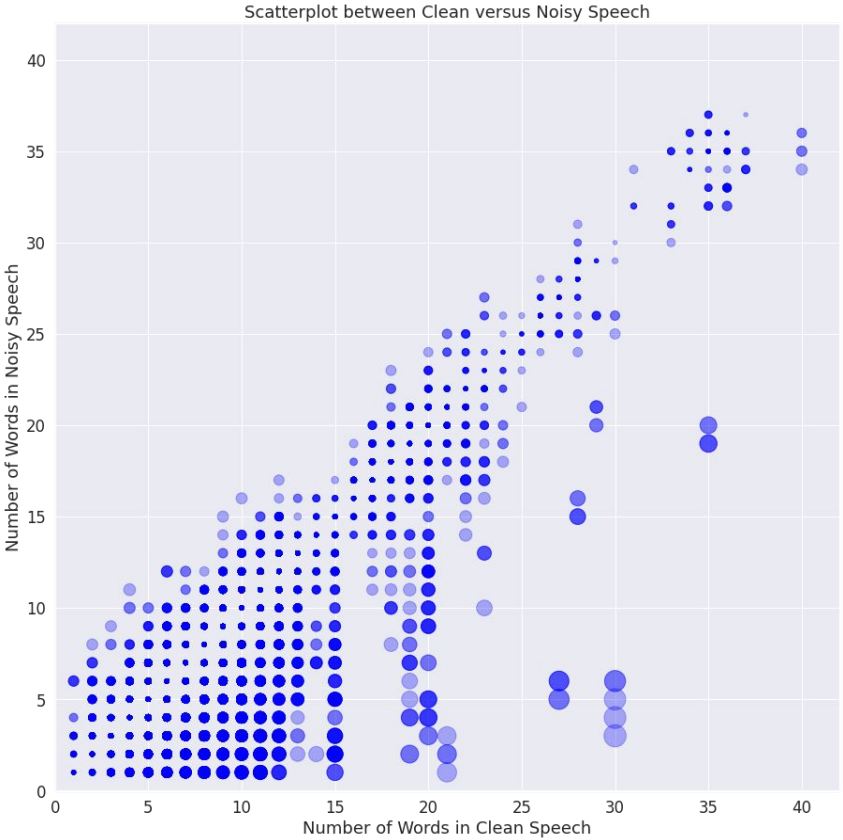
- Clean Audio Example 

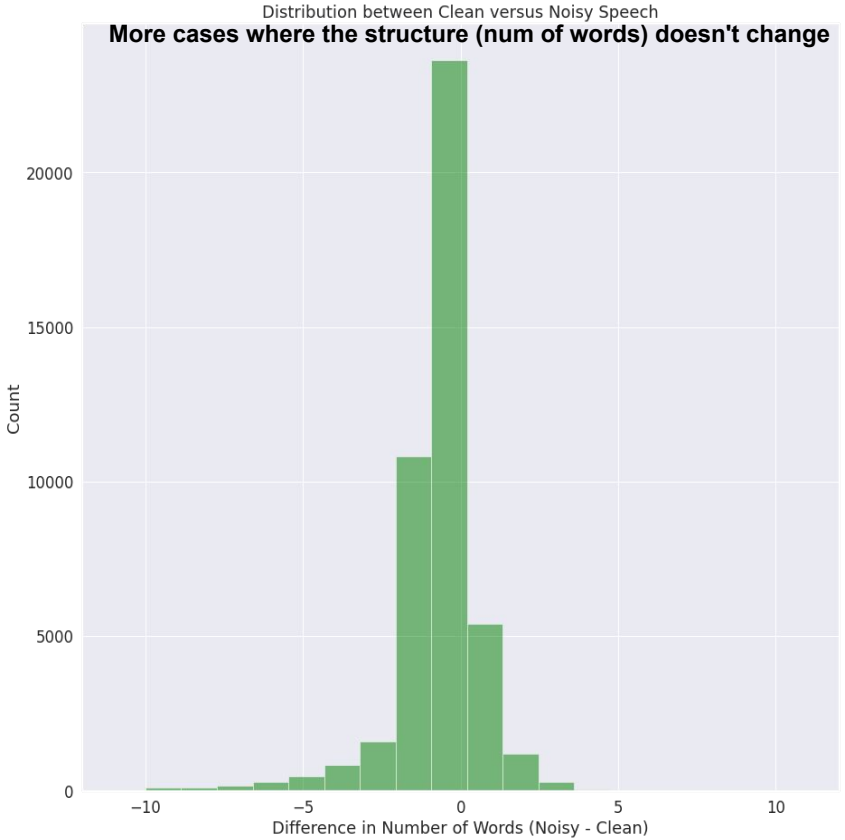
Speech to Text : `when the sunlight strikes raindrops in the absence of presidential in the rainbow`

- Noisy Audio Example 

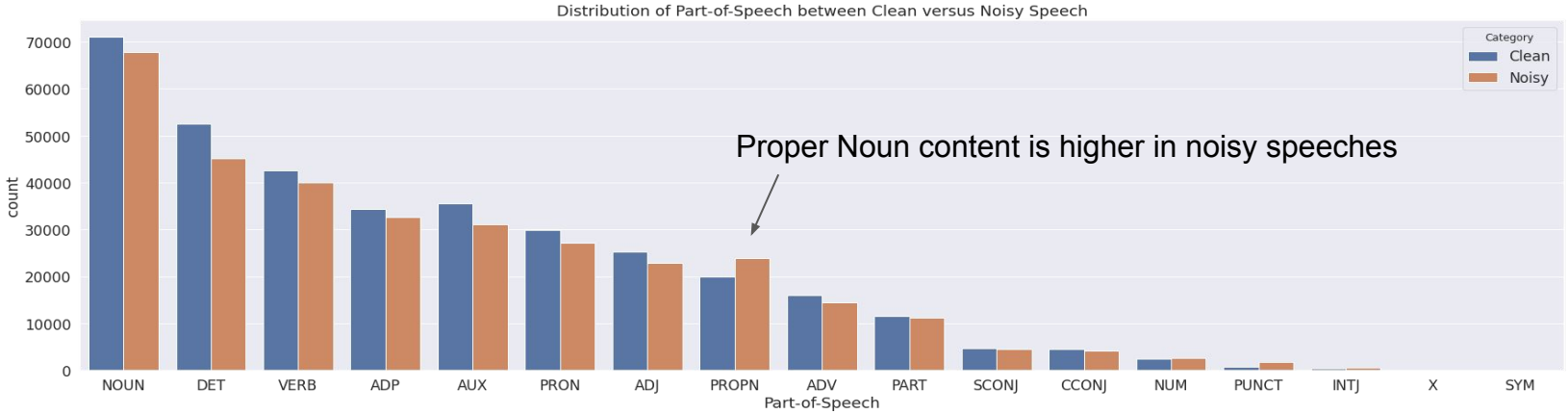
Speech to Text : `what is the Light Strike Crane Company`

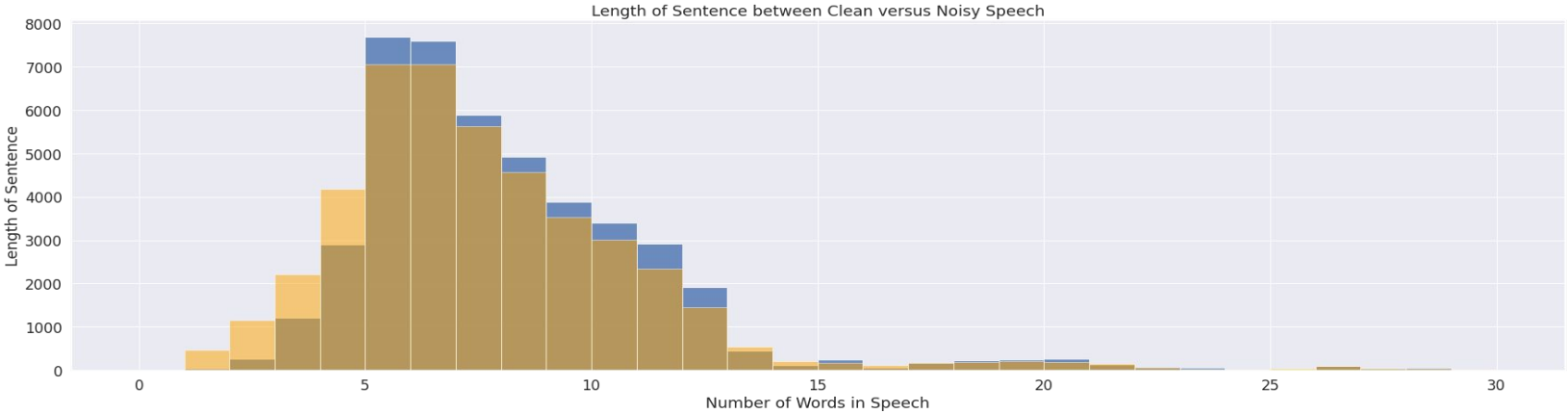
Speech Conversion Stats: Clean Audio vs Noisy Audio



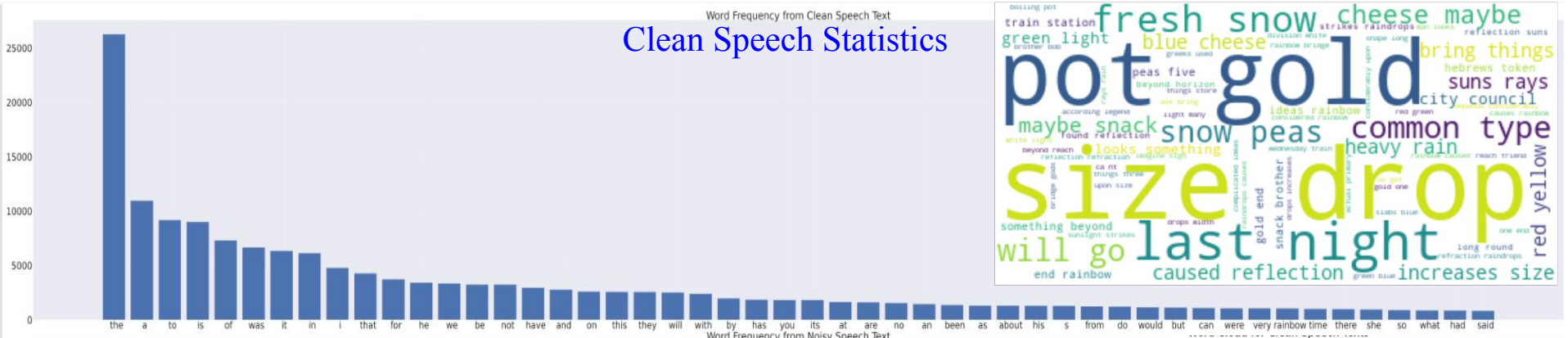


Speech Conversion Stats: Clean Audio vs Noisy Audio

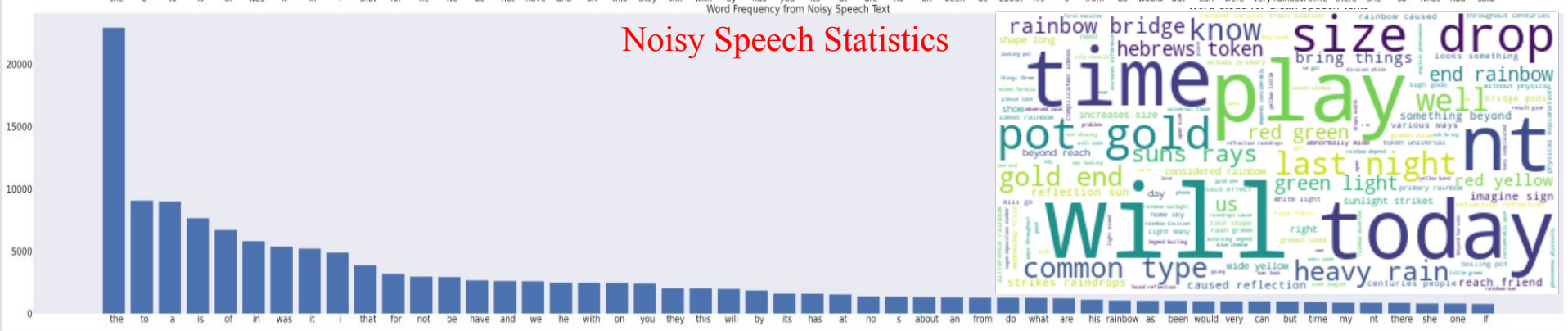




Speech Conversion Stats : Word Frequency

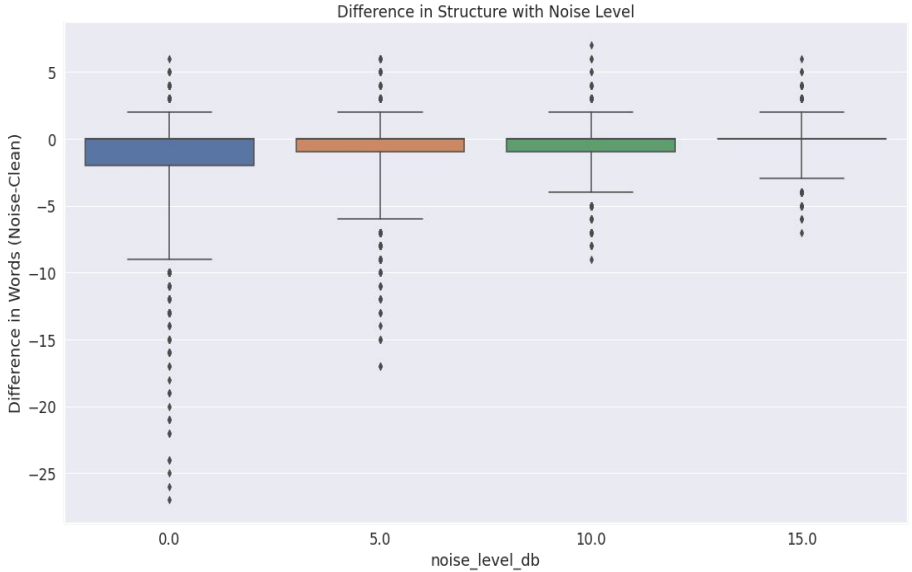


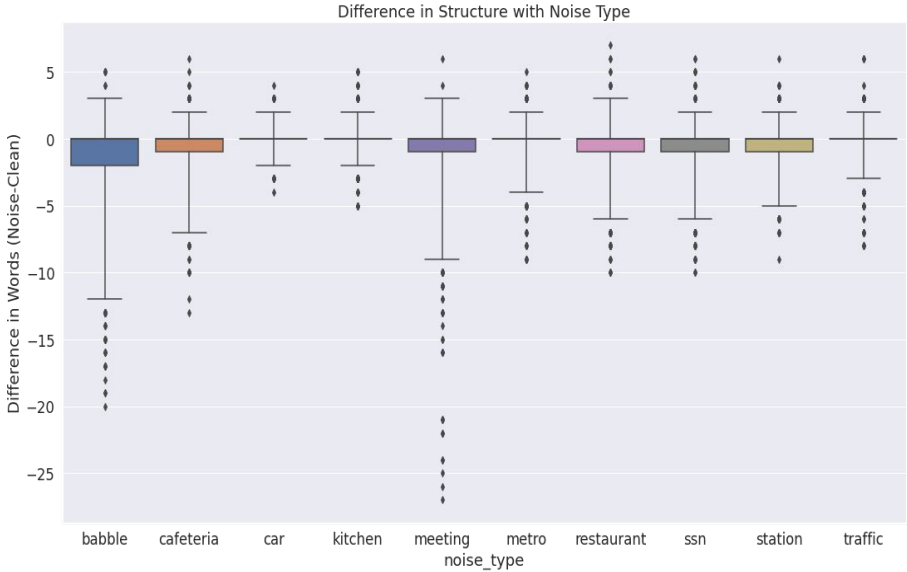
Clean Speech Statistics



Noisy Speech Statistics

Speech Conversion Stats: Effect of Noise





- Lower dB noise in speech generate more structure difference.
- Noise-type such as babble, cafeteria, meeting etc. give more variability in structures.

Observation: Speech Recognition Errors

- The presence of background noise affect the textual speech conversion from audio content.

Objective: Speech Error Prediction

- Construct a deep learning modelling using NLP to predict speech recognition errors generated as a result of background noise.

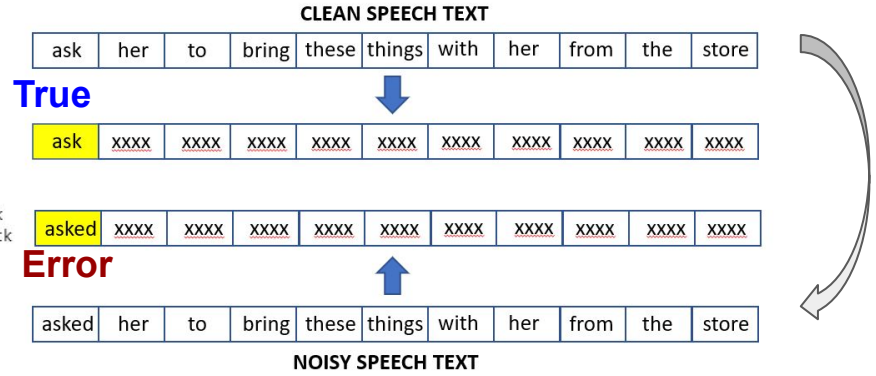
- In particular the model will predict **wording error** on speech translated texts.

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

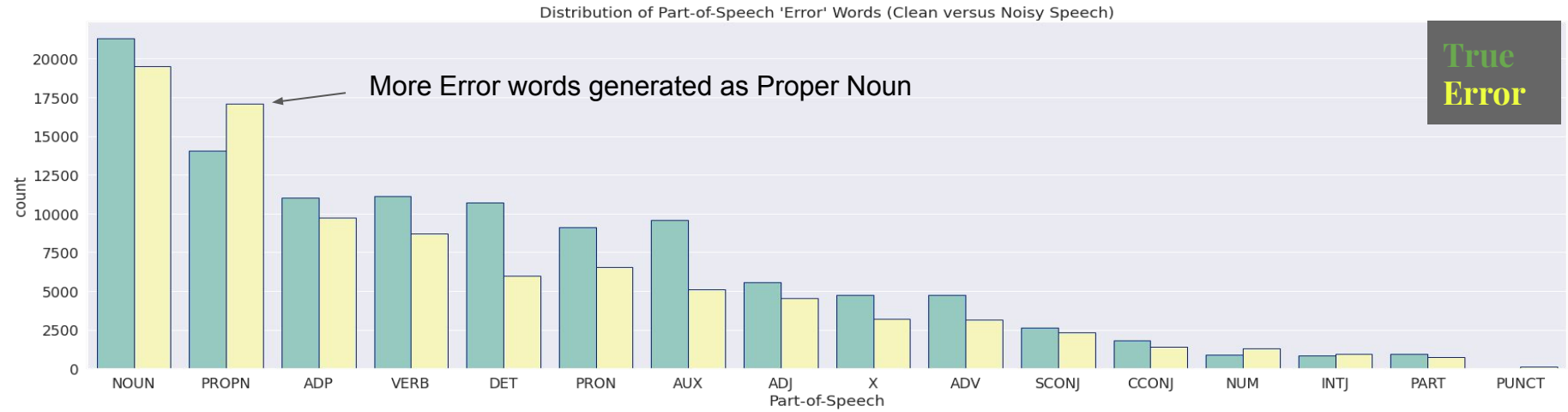
Data Transformation: Identify True vs Error Words

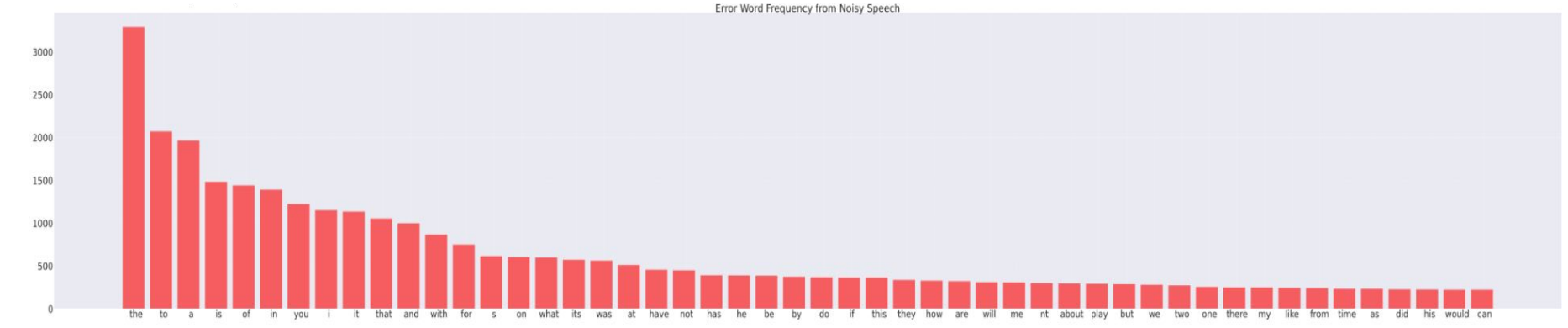
```
Cases where one word in clean was mis-translated to one word in noisy case

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 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

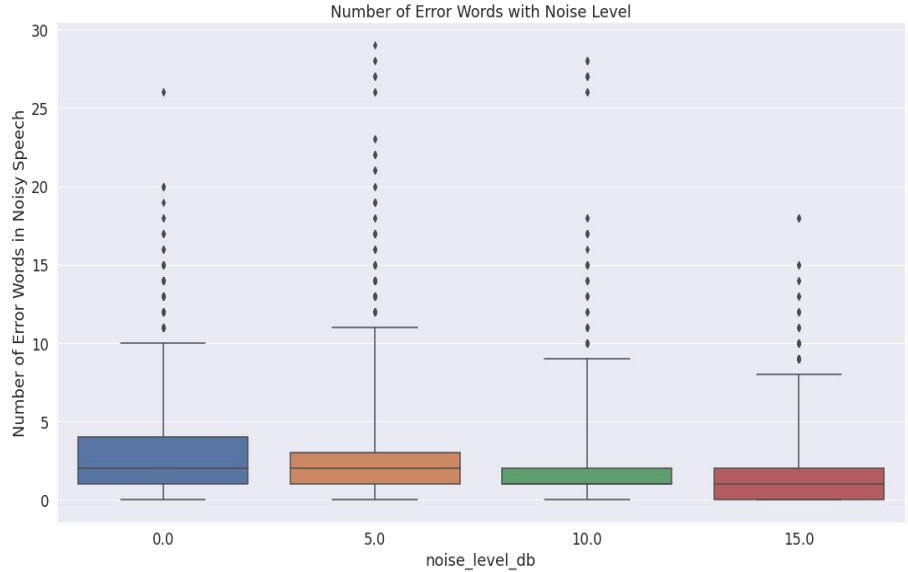


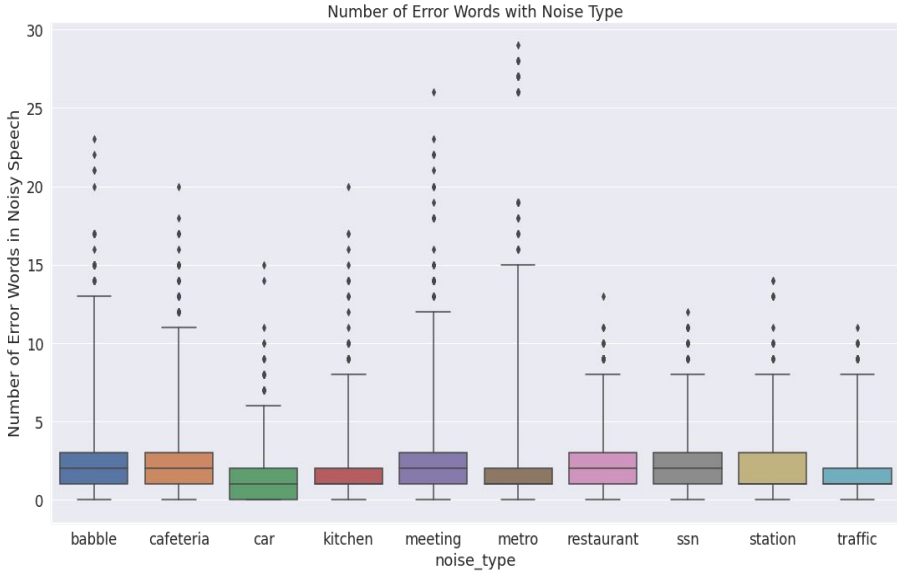
Data Exploration: True vs Error Words





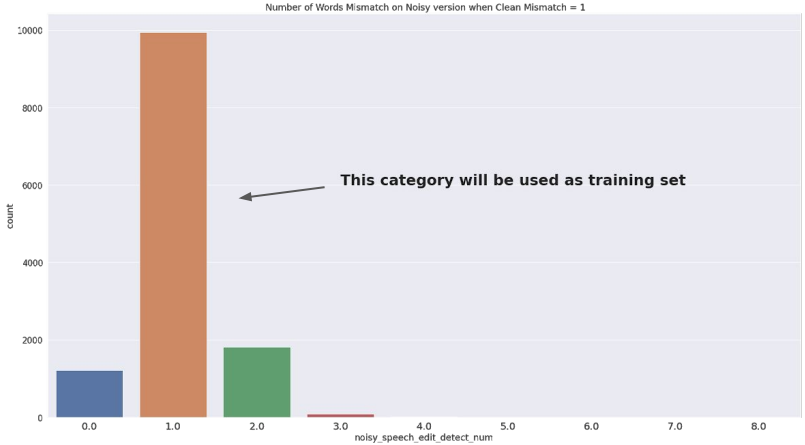
Data Exploration: True vs Error Words





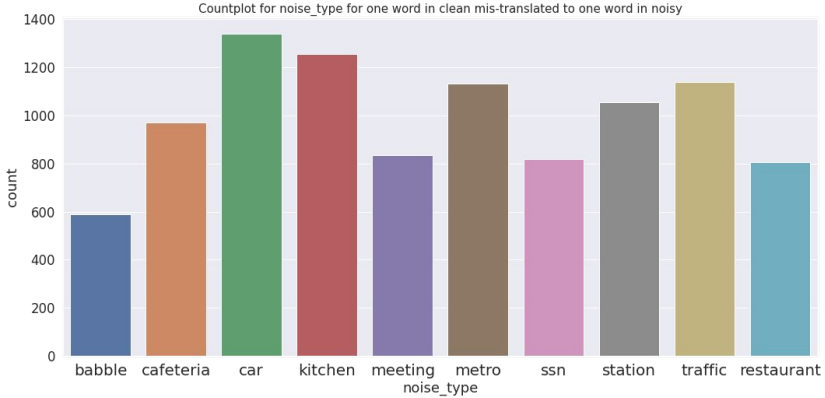
- Number of Error words are statistically more in cases where noise levels were low. (counter intuitive)

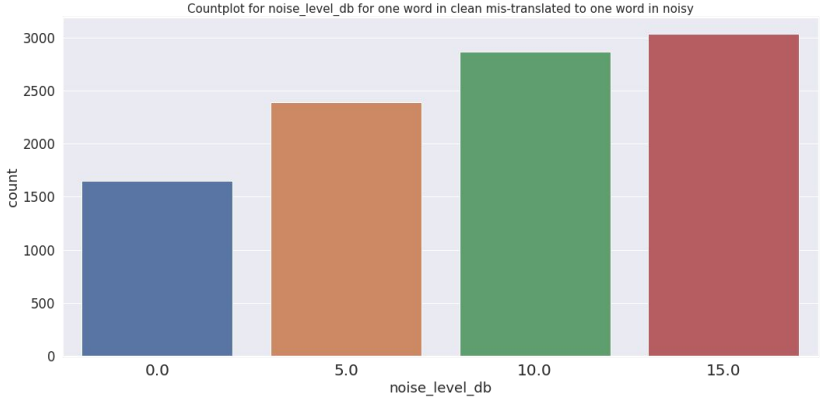
Feature Engineering: Data Selection



- Multiple words can be mistranslated as speech errors.

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





Data Transformation

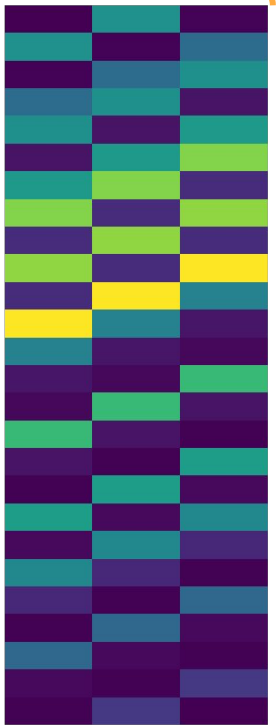
'when a man looks for something beyond his reach his friends say he was looking for the pot of gold at the end of the rainbow'

Error word

Input Sentence

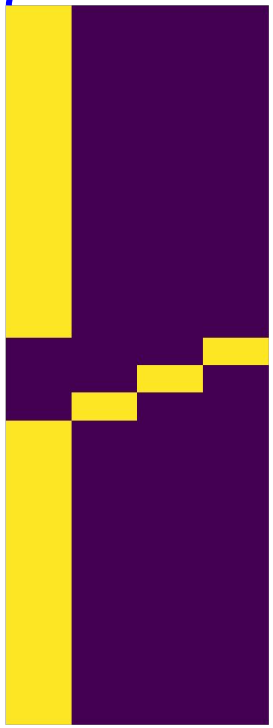
```
# when a
when a man
a man looks
man looks for
looks for something
for something beyond
something beyond his
beyond his reach
his reach his
reach his friends
his friends say
friends say he
say he was
he was looking
was looking for
looking for the
for the pot
the pot of
pot of gold
of gold at
gold at the
at the end
the end of
end of the
of the rainbow
the rainbow #
```

```
[  1. 125.   3.]
[125.   3.  88.]
[  3.  88. 124.]
[ 88. 124.  14.]
[124.  14. 134.]
[ 14. 134. 201.]
[134. 201.  32.]
[201.  32. 205.]
[ 32. 205.  32.]
[205.  32. 246.]
[ 32. 246. 110.]
[246. 110.  16.]
[110.  16.   6.]
[ 16.   6. 167.]
[  6. 167.  14.]
[167.  14.   2.]
[ 14.   2. 138.]
[  2. 138.   7.]
[138.   7. 115.]
[  7. 115.  29.]
[115.  29.   2.]
[29.  2. 84.]
[ 2. 84.  7.]
[84.  7.  2.]
[ 7.  2. 42.]
[ 2. 42.  1.]
```



DEEP LEARNING MODEL

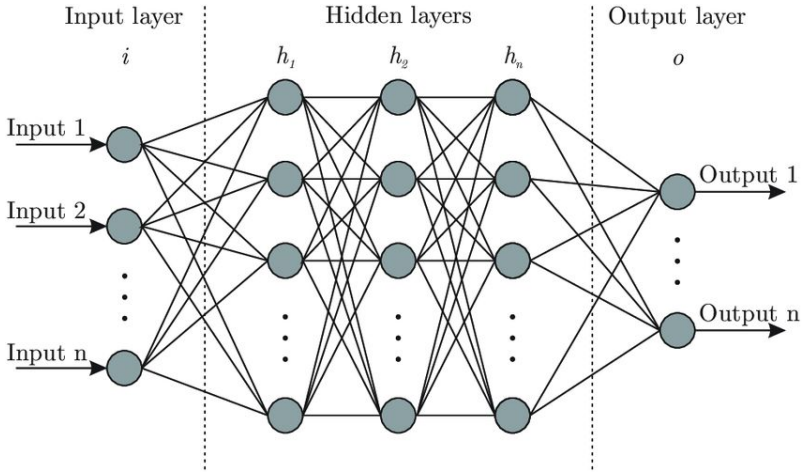
Output Error Word



```
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[0 0 0 1]
[0 0 1 0]
[0 1 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
[1 0 0 0]
```

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
```



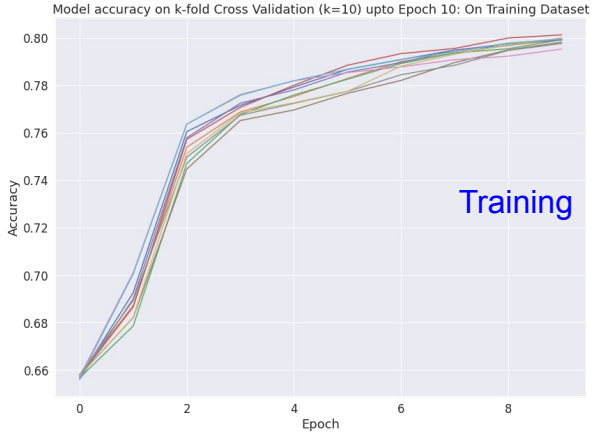
Error word : Probability Score

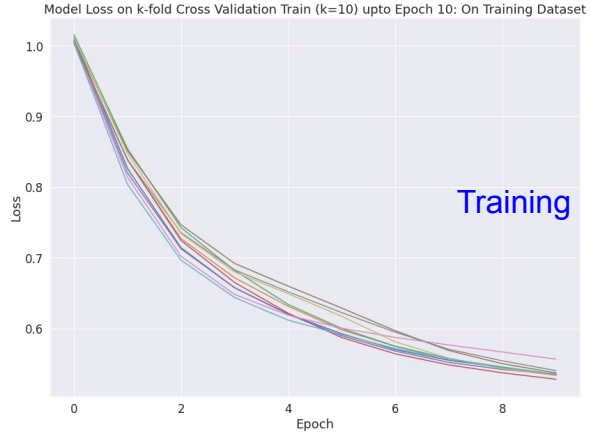
```
[{'beating': 0.67}]
[{'what': 0.67}]
['NO ERROR DETECTED']
[{'how': 1.0}]
[{'boe': 0.33}]
```

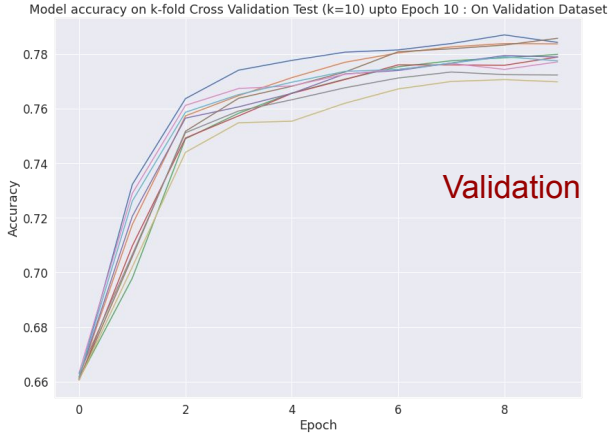
Modelling: K-Fold Performance

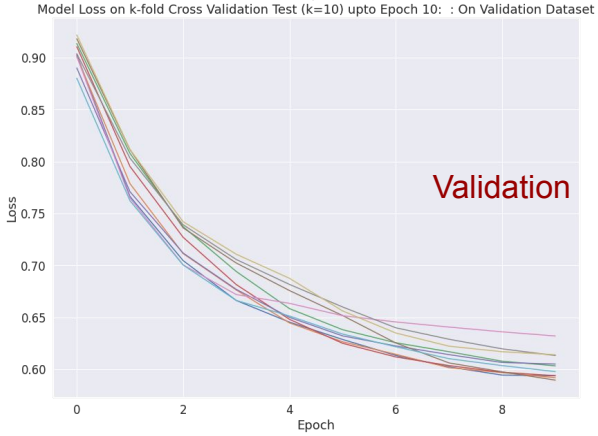
	validation accuracy	train accuracy
0	78.426254	81.173706
1	78.371328	81.278992
2	77.986819	81.021118
3	77.890688	81.584167
4	77.890688	81.201172
5	78.577316	81.004333
6	77.712166	80.586243
7	77.231532	80.847168
8	76.981187	81.338787
9	77.750307	81.038195

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

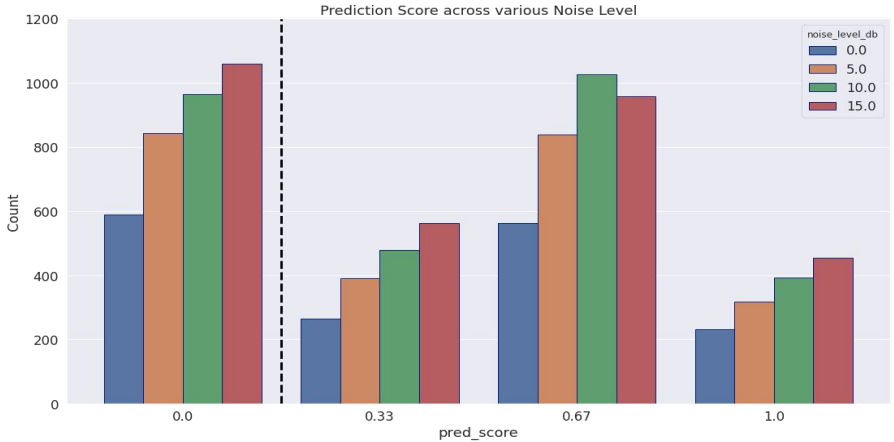




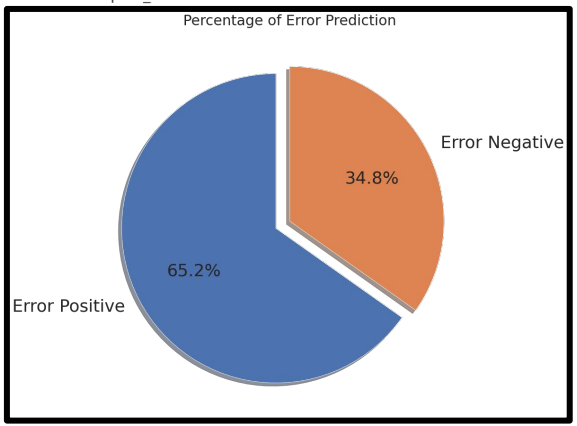




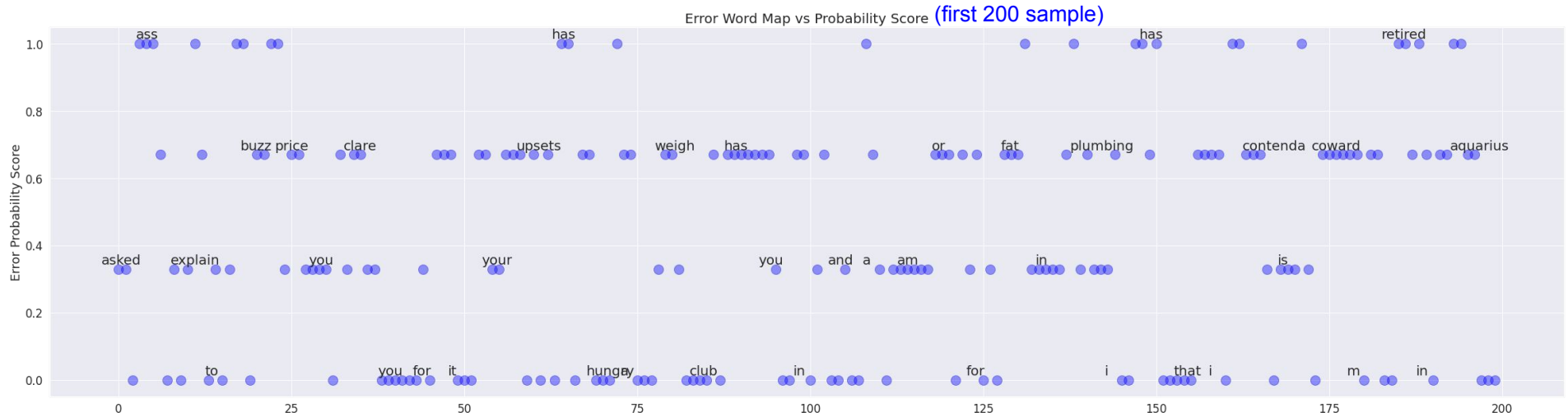
Modelling: Error Prediction Performance



- Prediction scores distribution have similar pattern across various noise levels and noise types.



Modelling: Word Map from Probability Score



- Some common pattern in words observed among various score group.

Modelling: Word Cloud from Probability Score

Word Cloud for Error Probability : 0

Word Cloud for Error Probability : 0.33

Word Cloud for Error Probability : 0.67

Word Cloud for Error Probability : 1

Modelling: Limitation

- Modelling predict errors in one wording only.
- Speech translation errors can also be made by other sources such as mispronunciations, language accent, voice quality etc.
- Training data is limited and need data augment to generalize the model.

Modelling: Future Directions

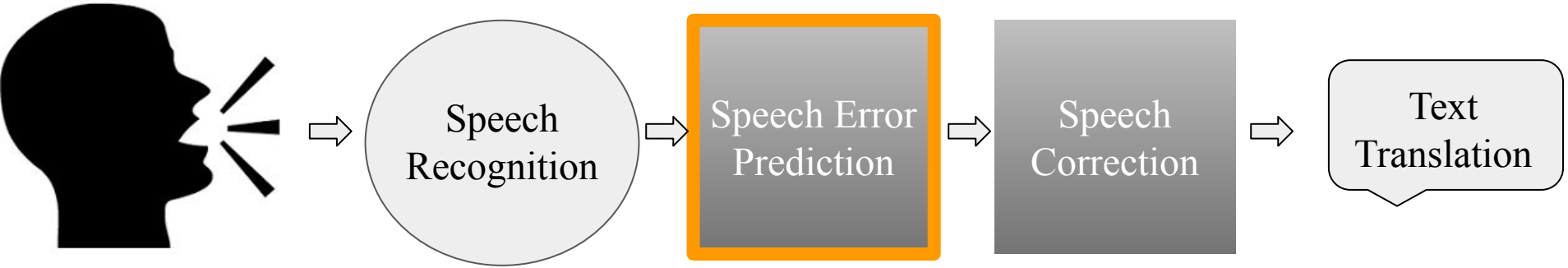
- Data augmentation by adding rhyming word to modelling inputs.

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

Implementation



Conclusion

- NLP and deep-learning model is shown to predict speech recognition error.

Questions ??

Feedback

- From Jay
  - Bilingual Errors
  - Dot Size
  - No Noise effect on prediction result
  - More frequent words have more errors
  - Different type of audio clips example