Contextualized Spoken Word Representations using Convolutional Autoencoders

Prakamya Mishra (1610110256)
Pranav Mathur (1610110257)

INTRODUCTION

- ★ This paper first depicts the process of splitting long audio files into individual words by cutting the silences from the file and calculating the Mahalanobis distance to identify these silences.
- ★ The second and main aim is to generate a deep learning model to visualise audio files in the form of unique vector representations.

RELATED WORK

- ★ Word2Vec has bestowed a prominent impact on the NLP field. With words being represented as vectors that provide context with reference to the nearby words, the vector representations can be used for sentiment analysis
- ★ There have been a few attempts by embedding word segments as fixed length vectors, which also proved to have a great application in Spoken Term Detection (STD).

WORD2VEC

- ★ Word2Vec is one of the most popular techniques to learn word embeddings using shallow neural network.
- ★ It's objective is to have words with similar context occupy close spatial positions in a graph.
- ★ There are two types of models available :

1. Common Bag Of Words (CBOW) Model
2. Skip-Gram Model:

AUDIO SPLITTING

- ★ Audio Splitting was achieved by using PyDub, a simple python library that can be used for audio manipulation at a high level interface.
- ★ This paper deals with the basic classification into the following three categories -

  (i) silence (S), where no speech is produced,

  (ii) unvoiced (U), where the vocal cords are not vibrating

  (iii) voiced (V), in which the vocal cords get tensed and allow air flow to generate speech.

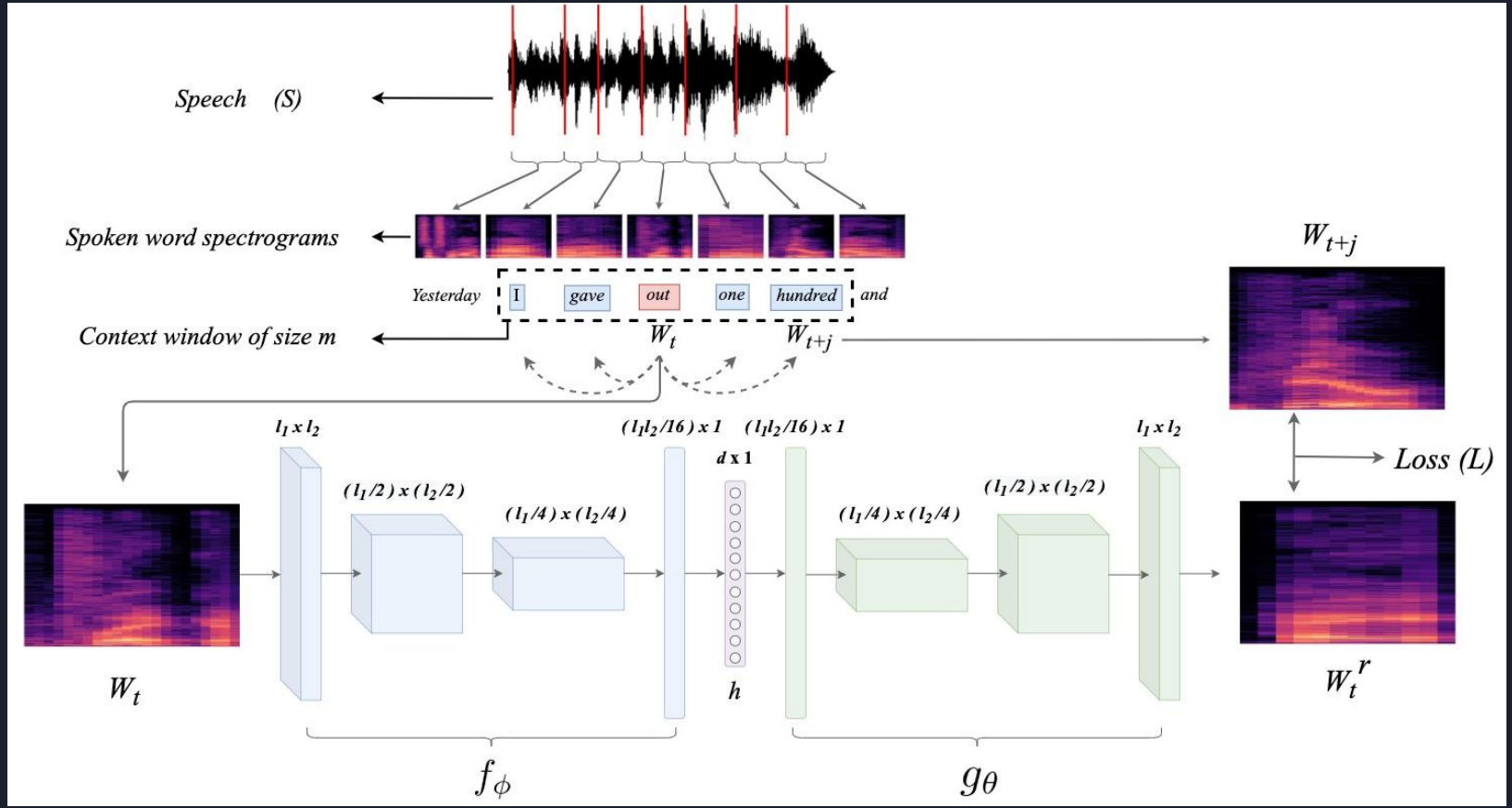
DATASET PREPARATION

- ★ The publicly available Trump speeches (Audio and word transcription) dataset [4] has been used for training the model.
- ★ This dataset contains the complete audio files as well as the JSON files that contain a list of all the words spoken in the speech.
- ★ A spectrogram is a visual representation of the spectrum of frequencies of an audio file as it varies with time. Each audio file was converted to an image spectrogram of size 1004 X 642 pixels.

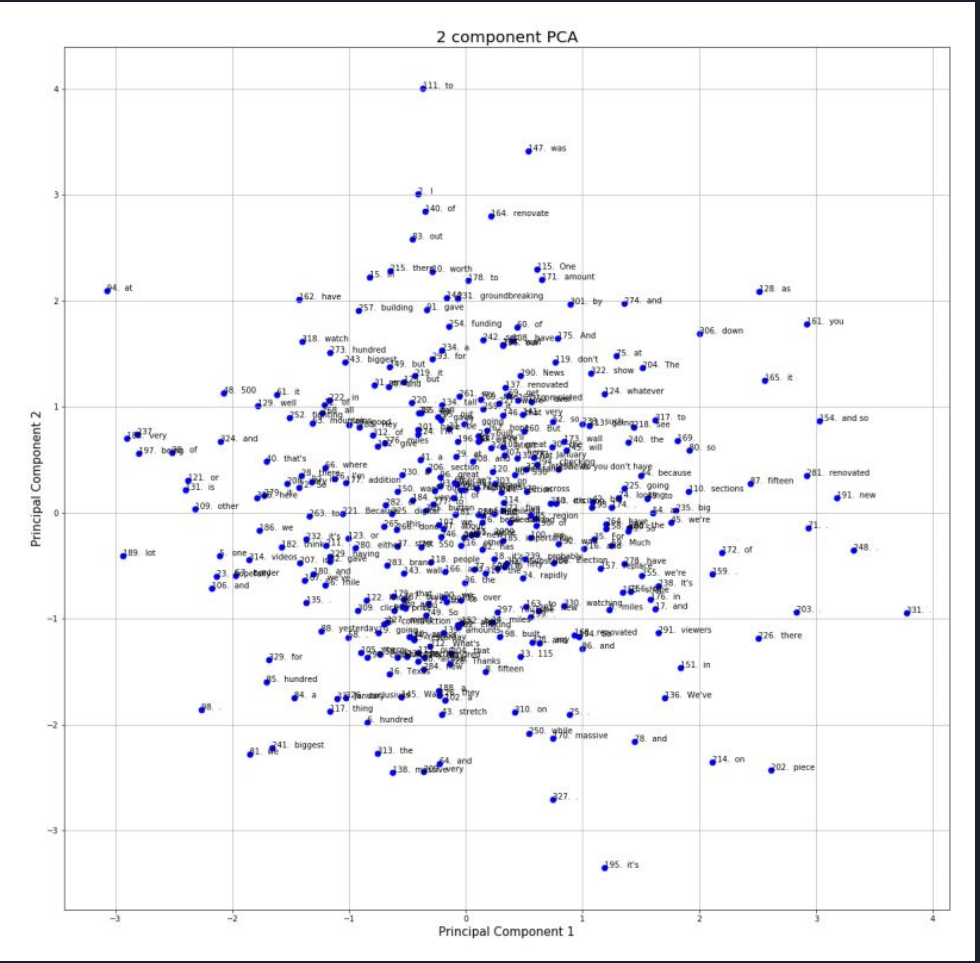
MODEL

- ★ **Convolutional Autoencoders have been used to learn a 16 dimensional latent vector representation of each spoken word using the context in which it was spoken.**
- ★ **This neural network model contains in itself 2 different models.**

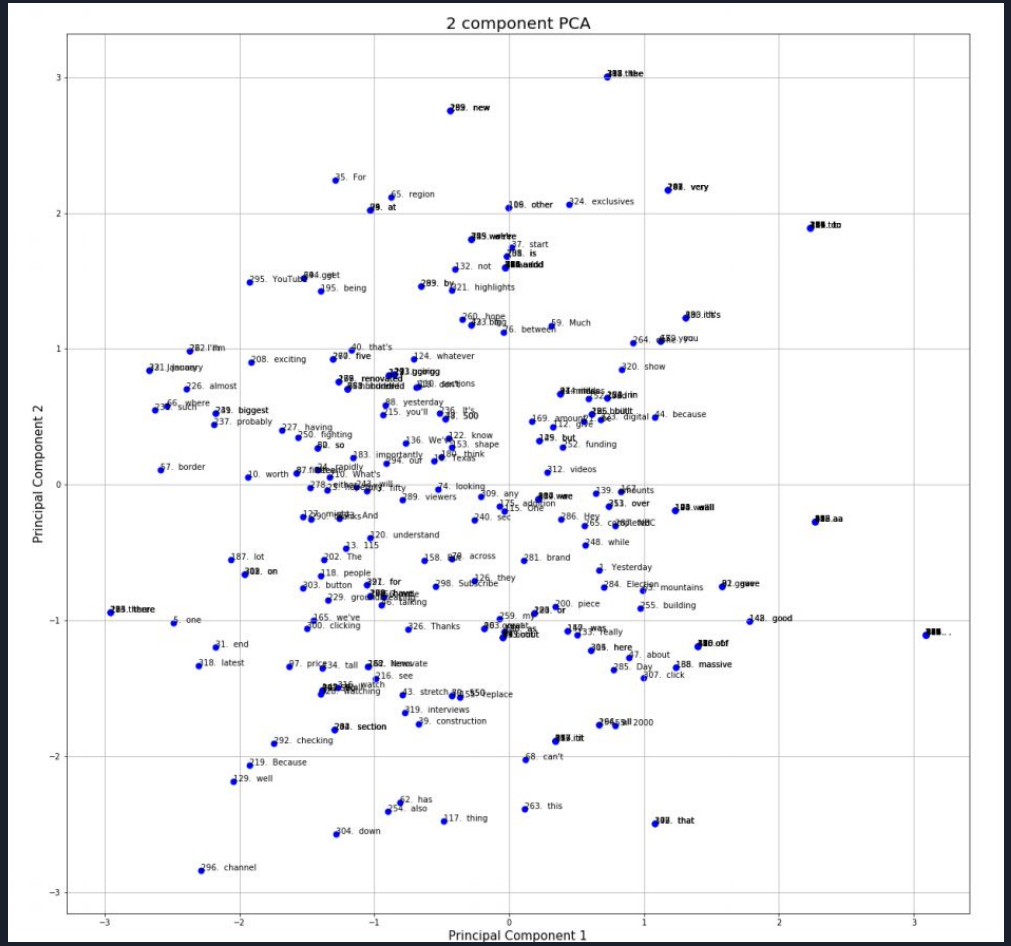
MODEL


Speech (S)
Spoken word spectrograms
Yesterday
I
gave
out
one
hundred
and
Context window of size m
W_t
W_{t+j}
W_{t+j}
Loss (L)
$l_1 \times l_2$
$(l_1/2) \times (l_2/2)$
$(l_1/4) \times (l_2/4)$
$(l_1 l_2/16) \times 1$
d x 1
$(l_1 l_2/16) \times 1$
$(l_1/4) \times (l_2/4)$
$(l_1/2) \times (l_2/2)$
$l_1 \times l_2$
W_t
h
W_t^r
f_ϕ
g_θ

Results


2 component PCA
Principal Component 2
Principal Component 1

Results


2 component PCA
Principal Component 1
Principal Component 2
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
35. For
45. region
79. at
109. other
124. exclusives
289. new
195. being
221. highlights
69. Much
64.
208. exciting
124. whatever
20. show
226. almost
37. probably
227. having
136. We
122. know
52. funding
57. border
10. worth
74. looking
89. viewers
20. understand
187. lot
202. The
158.
281. brand
1. Yesterday
18. people
203. button
298. Subscribe
284. Election
29. ground
200. piece
255. building
6. one
165. we've
60. clicking
226. Thanks
148. good
21. end
118. latest
97. price
34. tall
216. see
13. stretch
150.
replace
19. interviews
39. construction
92. checking
219. Because
129. well
68. can't
62. has
64. also
117. thing
263. this
304. down
296. channel
85. Day
107. click
88. massive
7. about
obf
2000