ASSR: Automatic Stuttered Speech Recognition

Anshul Gupta

IIT Bombay

November 30, 2017

Introduction

1. More than 70 million people worldwide are stutterers – that's one in every 100
2. State of the art speech-to-text systems fails miserably with accuracy as low as 18% and as high as 73% as compared to a baseline of 92% for normal speaker [4]
3. The existing work [1] that has been done for this problem is just classification of a speech as a stuttered speech or a normal speech

Dataset

1. University College London Archive of Stuttered Speech (UCLASS) [2] database
2. Recordings of monologues, reading and conversations of different speakers ranging from 7 to 20 years old
3. Most of them do not have time aligned labels and/or orthographic transcriptions
4. We are using 16 audio files have time aligned labels
5. .wav files having sampling rate of 22050Hz

Methodology

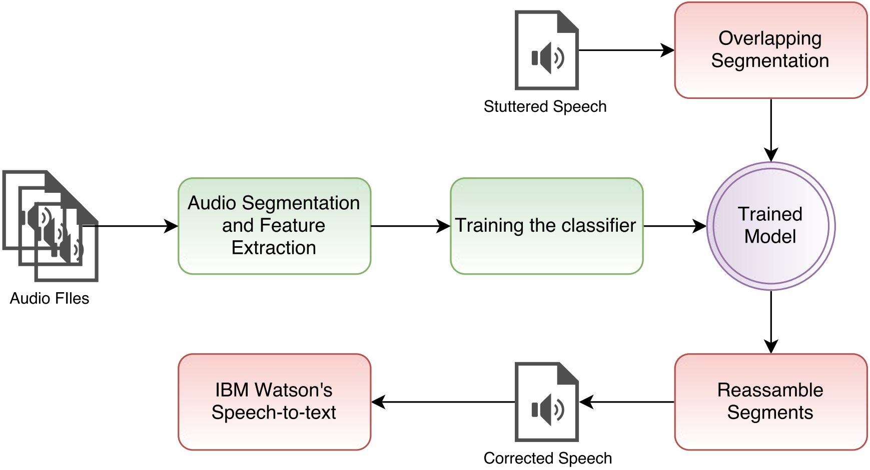

Figure: Flow Diagram

Methodology: Data Pre-processing I

Used the time-aligned transcriptions to split the data files into stuttered segments and normal segments

Table: Data Statistics

	ALL	STUTTER	NORMAL
COUNT	12633	2643	9990
MAX (ms)	17044	17044	14499
MIN (ms)	0	1	0
MEAN (ms)	315.0925	762.5323	196.7158
MEDIAN (ms)	192	486	168
MODE (ms)	109	201	93

Methodology: Data Pre-processing II

1. Very unlikely to see a 17 sec stuttered segment
2. We Segmented the files segments further down to less than or equal to 300 ms
3. This segmentation created 17,545 segments which were used for training the models

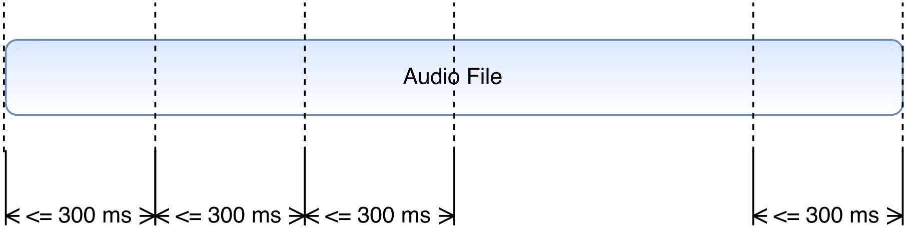

Figure: Audio segments

Methodology: Feature Extraction

1. MFCC features as they are very representative of human speech
2. RMSE which represents the loudness of a speech
3. Feature vector had mean and variance of $39 + 1$ features
4. Overall 80 features

Methodology: Classification

1. DNN gave the highest accuracy and took around ∼1 min to train as compared with SVC which took more than 1.5 hours.
2. DNN had 3 hidden layers, each having 10 neurons. Learning rate was 0.001 and training epochs were 1,200

Table: Classification Accuracy of models

	Accuracy (%)
DNN	87.07%
SVC	85.43%
Decision Trees	76.63%
Gaussian Naïve Bayes	76.63%
Bernoulli Naïve Bayes	71.43%
Multinomial Naïve Bayes	71.43%

Methodology: Audio Correction

With the classifier trained with an accuracy of $\sim$87%, next in the pipeline is audio correction.

Audio Correction: Overlapping Segmentation I

1. Model is trained on audio segments of duration $<=$ 300ms
2. It was only obvious that the audio to be corrected needs to be segmented with duration of 300ms
3. Less obvious was to detect the stutter boundaries
4. Instead of naïvely segmenting the audio in contiguous manner, we overlapped the segments
5. We could detect the stuttered and non-stuttered parts with the granularity of 100ms

Audio Correction: Overlapping Segmentation II

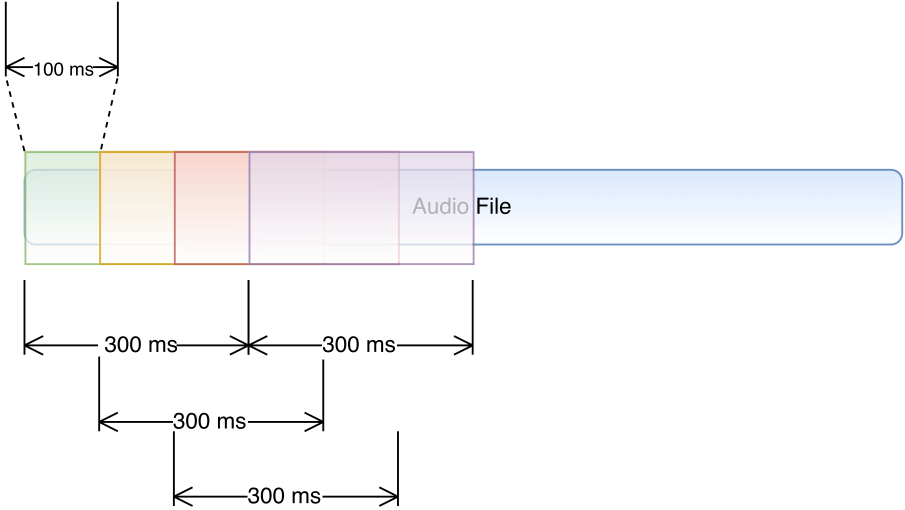

Figure: Overlapping Segmentation

Audio Correction: Re-assembling the segments I

1. Classifier gave the labels of the overlapping segments
2. Remove the segments which were labelled as STUTTER and combine the segments labelled as NORMAL
3. One way of assembling the segments was to append contiguous chunks together
    1. This will result in sharp interjections at the point of concatenation
    2. Very artificial sounding voice
4. So instead of naïvely appending the adjacent chunks, we interpolated the audio samples between the end of the previous chunk and the beginning of the current chunk

Audio Correction: Re-assembling the segments II

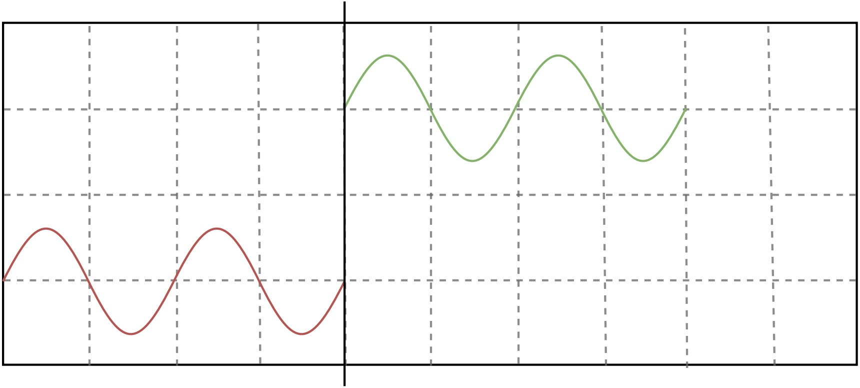

Figure: Naïve Re-assembling

Audio Correction: Re-assembling the segments III

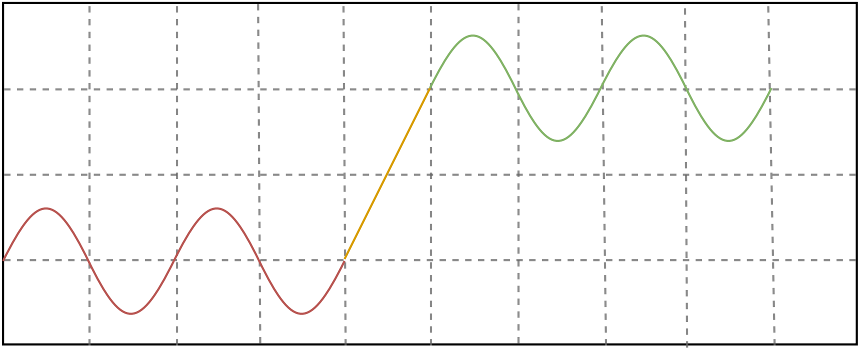

Figure: Smoothed Re-assembling

Speech-to-text

1. The UCLASS dataset [2] is in British English
2. We used the IBM Watson's Speech-to-text [3] which already has a trained model for GB English.

Results

1. As we can see, our model is far from perfect
2. But we achieved some improvement over the un-modified audio files
3. With the limited dataset available for training, we could only achieve this much accuracy

Table: Comparison of WER of original and the corresponding corrected audio

Subject	Original (%WER)	Corrected (%WER)
M_0017_19y2m_1	74.928%	73.775%
M_0065_20y1m_1	125.000%	116.429%
M_0100_12y3m_1	84.173%	89.928%
M_1017_11y8m_1	55.396%	48.921%
M_1017_13y2m_1	59.322%	46.610%

The source code of the project can be found at https://github.com/anshulgupta0803/stutter-speech-recoginition

References

Manu Chopra, Kevin Khieu, and Thomas Liu.
Classification and recognition of stuttered speech.

Peter Howell, Stephen Davis, and Jon Bartrip.
The university college london archive of stuttered speech (uclass).
Journal of Speech, Language, and Hearing Research, 52(2):556–569, 2009.

IBM.
Ibm watson speech to text.
https://www.ibm.com/watson/services/speech-to-text.

Emily Mullin.
Why siri won't listen to millions of people with disabilities, May 2016.
Article.