Classification of Stuttered Speech Behaviors of Filipino People Who Stutter Using a Multilayer Perceptron Neural Network

Project Team Members

Juan Diego V. Huet

Aaron Andrei D. Ambas

Kyle Rafael F. Sulabo

I. Introduction

Background of the Problem

Stuttering is "…a speech disorder in which the flow of speech is disrupted by involuntary repetitions and prolongations of sounds, syllables, words or phrases as well as involuntary silent pauses or blocks in which the person who stutters is unable to produce sounds". (World Health Organization, 2010). The primary behaviours manifested by stuttering are repetition, prolongation of sounds, and blocking (Northern Arizona University, n.d.). A repetition occurs whenever the person who stutters repeats a sound, syllable, or a one syllable word more than once or twice; a prolongation occurs when a speech sound is held out but the mouth, lips, or tongue stops moving; blocks occur when the person stops the flow of sound or air in the lungs, throat, mouth, lips, or tongue. (HomeSpeechHome.com, n.d.)

A speech language pathologist conducts speech therapy as well as diagnosis. Speech therapy is the process of correcting disfluency in speech. A speech diagnosis is needed before the actual speech therapy in order to accurately decide what treatment a patient should receive (Scott, 2008).

The speech pathologist conducts the diagnosis by first holding a casual conversation with the patient. Behaviors that are manifested physically by the patient are then documented by the speech pathologist. Afterwards, tests are conducted to measure the severity of stuttering by conducting speech focused tests that will test the patient's fluency. The patient will be asked to read a passage where the percentage of the total number of words stuttered over the total number of words in the passage will be recorded. The patient will also be asked to hold a conversation with the speech pathologist with the patient discussing ideas about a certain topic, this time the patient's voice will be recorded for use as reference. The percentage of the total number of words stuttered over the total number of words used by the patient will also be recorded. The speech pathologist also takes note of the different types of disfluency observed while conducting the speech tests.

The traditional method of diagnosing speech is generally time consuming. Another issue with it is that different speech pathologists may make their own different judgements when diagnosing (Kully and Boerg, 1988).

Statement of the Problem

- How can stuttered speech behaviors be classified more objectively with the use of a Multilayer Perceptron model?
- How effective will a Multilayer Perceptron be in classifying stuttered speech behaviors?

Objectives

- Implement a Multilayer Perceptron classifier to classify audio with stuttered speech instances based on the extracted audio features.
- Measure the effectiveness of the proposed MLP classifier model.

Significance

The findings of this research will benefit the following agencies:

To People who Stutter

This project will benefit patients under speech therapy that will undergo speech diagnostics. Using a classifier model to classify repetitions and prolongations can make the process of differentiating repetitions from prolongations more objective. Having a more objective classification of stuttered speech behaviors can improve the accuracy of diagnoses.

To Speech Pathologists

Speech Pathologists will be able to use this project to produce more accurate and more objective results when conducting a speech diagnosis. Better diagnostic

results will lead to more appropriate therapies among patients. This can also be used alongside the traditional method of diagnosis to lessen ambiguity.

To Other Researchers

Other researchers will benefit from this study by having a better idea of what machine learning algorithms will be most effective when classifying audio features in the context of stuttered speech.

II. Scope and Limitations

The study aims to measure the effectiveness of a Multilayer Perceptron Neural Network in classifying stuttered speech. The research will only focus on the classification of segmented stuttered speech instances. Explicitly, the research does not aim to detect stuttered speech from speech audio. The expected input of the classifier models are audio features from segments of stuttered speech sounds, namely repetitions and prolongations. Furthermore, the project does not also aim to implement the results of the research into a working product, but rather aims to prove the proposition of whether or not multilayer perceptrons can be used to classify stuttered speech behaviors.

III. *Related Literature*

Behaviors of Stuttering

Stuttering behaviors are mainly characterized into primary and secondary. The primary behaviors are:

- Repetitions of sounds, syllables and words
- Prolongation of single sounds
- Blocks of airflow when speaking

Secondary behaviors include:

- Hesitations
- Interjections of sounds, syllables of words (ahh, uhm)
- Word revision, word changes

- Unnecessary motor movements

(Northern Arizona University, n.d.)

Related Researches

There have been numerous researches conducted that proposed to classify or identify stuttered speech with the use of machine learning, psychology, linguistics, and digital signal processing. The following is a discussion of related researches that the researchers have found relevant.

In 1995, Howell and Sackin classified stuttered speech into repetitions and prolongations with the use of an artificial neural network(ANN). They extracted 39 acoustic parameters, 20 vector based on autocorrelation function plus spectral coefficient based on a 19 channel vocoder. Envelope of speech waveform was obtained by filtering the signal using a 10hz lowpass filter. The best hit/miss rate was 0.82 for prolongations and 0.77 for repetitions. Further research was done by Howell, Sackin, and Glenn in 1997. In this research, they explored classification of stuttered words from fluent words, as well as classify stuttered words into prolongations and repetitions using ANN. For the dataset, they employed 12 children who speak stuttered English. The speech samples can be obtained from the University College London Archive of Stuttered Speech (UCLASS). For this study, speech were manually segmented into individual words. For the attributes, the researchers used whole word and part word duration; whole word, first part, and second part fragmentation; whole word, first part, and second part spectral measure; and part word energy. These attributes are then input to the networks. The classifier yielded 95% accuracy for fluent words and 78% for disfluent words. From the disfluent words, a 58% accuracy was achieved for prolongations and 43% accuracy for repetitions.

Geetha et. al researched on using ANN to classify children who stutter from children with normal non-fluency by using medical data as input. Fifty-one children were employed as respondents; the medical data of 25 children were used to train the ANN, and 26 were used for testing. The attributes that are used as input are age, sex, type of disfluency, frequency of disfluency, duration, physical concomitant, rate of speech, historical, attitudinal and behavioral

scores, and family history. A 92% accuracy of predicting normal non-fluency and stuttering was achieved.

Szczurowska et al. experimented with using Konohen and Multilayer Perceptron networks in classifying fluent and disfluent speech. Recordings were taken 8 stuttering Polish speakers and segmented disfluent 4-second-long fragments containing disfluency. Speech of fluent speakers containing the same fragments were also recorded. All utterances were analysed by FFT 512 with the use of a 21 digital 1/3-octave filters of centre frequencies between 100 and 10000 Hz and an A-weighting filter. FFT time resolution was 23 ms, which transformed every 4-second sample into 21 vectors consisting of 171 time points. These were then inputted to different types of MLP networks for training. A best accuracy of 76.67% was achieved.

A study by Ravikumar et al. explored the possibility of automatic detection of repetitions in read speech. This study proposed automatic detection of repeated syllables using 4 major steps they defined: segmentation, feature extraction, score matching, and decision logic. During the data gathering part, the researchers employed 10 people who stutter with a mean age of 25 as the respondents. A standard English Passage of 150 words was selected as the reading passage to be read by the respondents and these speeches were recorded at a sampling rate of 16000 samples per second. The collected audio samples are first manually segmented by the researchers into syllables. After segmentation, the segmented speech syllables are subject to feature extraction; 12 Mel Frequency Cepstral Coefficients(MFCC) were extracted from the segmented speech syllables. Score matching is then done using the Dynamic Time Warping Algorithm. The angle between between the segmented speech syllables were computed by using DTW on the MFCC's. These values were given to the decision logic to identify whether the syllables were repeated or not. The perceptron algorithm was used as the decision logic. This proposed approach by the researchers achieved an accuracy of 83%.

Another study by Mahesha and Vinod proposed classification of disfluent speech using k-Nearest Neighbors algorithm and Support Vector Machine. The researchers defined disfluent speech as speech containing repetitions, prolongations, interjections, and pauses. For the data gathering, audio samples used in the study was obtained from the University College London's Archive of Stuttered Speech (UCLASS). Samples are taken from standard reading of 25 different speakers with age between 10 to 20 years. After obtaining the samples, the

researchers segmented disfluent speech manually. Another pool of fluent speech is produced by 20 fluent speakers with a mean age group of 25 using the same reading passages used in the UCLASS database. Fluent and disfluent speech were segmented manually from these speech samples; overall, a speech corpus that contains 50 fluent and 50 disfluent speech segments is created. 40 fluent and 40 disfluent speech segments will be used for training the models, while 10 fluent and 10 disfluent speech segments will be used to test the models. Next, the MFCC's are obtained from the fluent and disfluent speech samples and used as the data to train two models with two machine learning algorithms (K-NN classifier and Support Vector Machine). The produced model using k-nn classifier achieved an accuracy of 86.67% in classifying disfluent speech and 93.34% in classifying fluent speech while the produced model using support vector machine algorithm achieved an accuracy of 90% in classifying disfluent speech and 96.67% in classifying fluent speech.

Hariharan et al. proposed classification of speech disfluencies with MFCC and LPCC features. The study used speech samples from UCLASS, speech samples from 39 people were used for this study; it includes one sample each from 2 female speakers and 37 male speakers ranged between 11 years 2 months and 20 years and 1 month. The speech samples contain speeches of reading passages specifically "One More Week to Easter" and "Arthur the Rat"; each of the two passages contains more than 300 words. The two types of disfluencies considered by the researchers were repetitions and prolongations and were identified and segmented manually by listening to the recorded speech signals. The segmented speech samples were downsampled from 44.1khz to 16khz and pre-emphasized with a high pass filter. After this the MFCC's and LPCC's were extracted from the segmented speech samples with different frame lengths and overlaps. The study experimented with 10, 20, 30, 40, and 50ms frame lengths and no overlap, 33.33%, 50%, and 75% overlap. The study also experimented with different high pass filters (alpha=0.91 to alpha=0.99). After feature extraction, k-Nearest Neighbor and Linear Discriminant Analysis classifiers were used. The best results achieved for MFCC and LPCC features are: 94.51% for LPCC (frame length=30ms, window overlap=75%, alpha=0.98) and 92.55% for MFCC (frame length=20ms, overlap=50%, alpha=0.9375).

Based on the literature review conducted, no known research has been done yet that explores the possibility of using a Multilayer Perceptron(MLP) model to classify extracted MFCC's from stuttered speech into repetitions or prolongations. Moreover, no known research

has been yet conducted that utilizes Filipino stuttered speech as the dataset. The researchers aim to fill this gap by training and testing a MLP model to classify stuttered speech from Filipino people who stutter as the experiment and reporting on the results of the effectiveness of the said classifier in this paper. The researchers' motivation to use a Multilayer Perceptron classifier to classify MFCC's from stuttered speech stems from the findings of Nanopoulos et. al. In their research, the use of a Multilayer Perceptron Neural Network on time-series data was explored.

Tools Used

jAudio

jAudio is a framework for audio feature extraction. The system meets the needs of Music Information Retrieval(MIR) researchers by providing a library of analysis algorithms that are suitable for a wide array of MIR tasks. In this research, jAudio will be used to extract the needed features that will be fed to the MLP network.

Weka

Weka is a collection of machine learning algorithms written in Java and developed at the University of Waikato, New Zealand. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. In this study, the researchers are going to use Weka to implement the Multilayer Perceptron classifier.

III. ***Theoretical Background***

Machine Learning

Machine Learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed (whatis.techtarget.com, nd.). It involves development of computer programs that can teach themselves to grow and adapt to change when new data is exposed. Machine Learning systems look through data to look for patterns and detect those patterns in the data to and adjust actions accordingly. A machine learning model refers to the model artifact created by training (Amazon Web Services Documentation, Amazon Machine Learning Guide, Training Machine Learning Models).

Machine Learning algorithms are usually classified into two: Supervised Learning and Unsupervised Learning. Supervised learning uses what it has learned in the past and applies it to new data. Some examples of algorithms that belong to Supervised Learning are Support Vector Machines, linear regression, naive Bayes, Neural Networks, etc. Unsupervised Learning on the other hand describes hidden patterns from unlabeled data. Some examples are clustering through k-means, mixture models, or hierarchical clustering; anomaly detection; and Neural Networks.

Mel Frequency Cepstral Coefficient

Introduced by Davis and Mermelstein in the 1980's, MFCC represents the envelope of the short time power spectrum manifested by the shape of the vocal tract of humans. MFCCs give accurate representations of phonemes being produced (practicalcryptography.com, n.d.).

Multilayer Perceptron

IV. Proposed Solution to the Problem (Methodology)

The researchers proposes the use of Mel Frequency Cepstral Coefficients as inputs to a Multilayer Perceptron Neural Network to classify repetitions and prolongations.

A. Segmentation

Audio samples of stuttered speech are obtained from 4 different Filipino stutterers. Currently, this is the pool of samples where the researchers will get the segmented stuttered speech. Only stuttered instances are to be segmented from the different audio files manually. Other related researches have segmented the stuttered speech by instance, such as P.

Mahesha and D.S. Vinod's research (An Approach for Classification of Disfluent and Fluent Speech using K-NN and SVM). After segmenting the stuttered instances, these instances will be classified into repetitions or prolongations. Currently, there are 46 audio clips of stuttered instances of which 22 of them are repetitions, and 33 are prolongations.

B. Feature Extraction

The segmented stuttered instance clips will each be then analyzed by the program jAudio. Two audio features will be extracted, namely the Mel Frequency Cepstral Coefficients and the Linear Predictive Coefficients for each frame. The two algorithms are performed by jAudio in audio feature extraction. The results will be manually classified whether the the results belonged to a repetition or prolongation. The MFCC's and LPCC's per frame will form a single vector and will be zero-padded to match the vector of the instance with the most number of frames. This will result into the instances having equal length of vectors.

C. Machine Training

The feature vectors will then be inputted to the MLP neural network. The number of neurons in the input layer will depend on length of the vectors, and the number of neurons in the output layer will be 2, taking into consideration the two classes: repetitions and prolongations. Different learning rates and momentum for will be tested. The result will be MLP models that will classify input MFCC's or LPC's into whether it belongs to a repetition or a prolongation.

D. Machine Testing and Analysis of Results

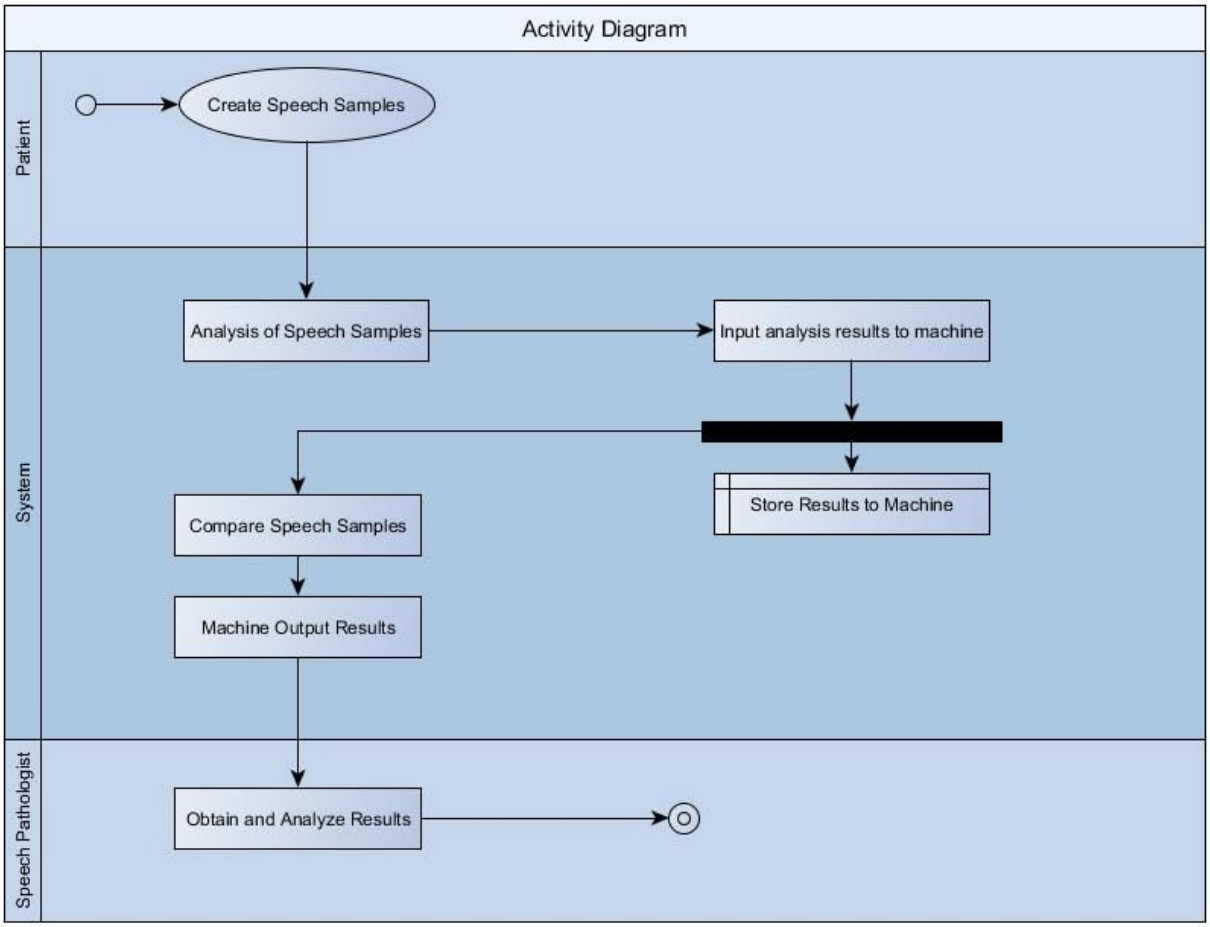
For testing, 10 fold cross-validation will be used. This means that 10 separate models will be built using 90% of the dataset, and 10% will be used to test the dataset. The testing set will be different for each model built, and the accuracy will depend on the combined accuracy of all the models built. The extracted audio features will be analyzed by the models and classify

them according to which behavior they might be. The accuracy of each model will be analyzed and discussed in the paper.

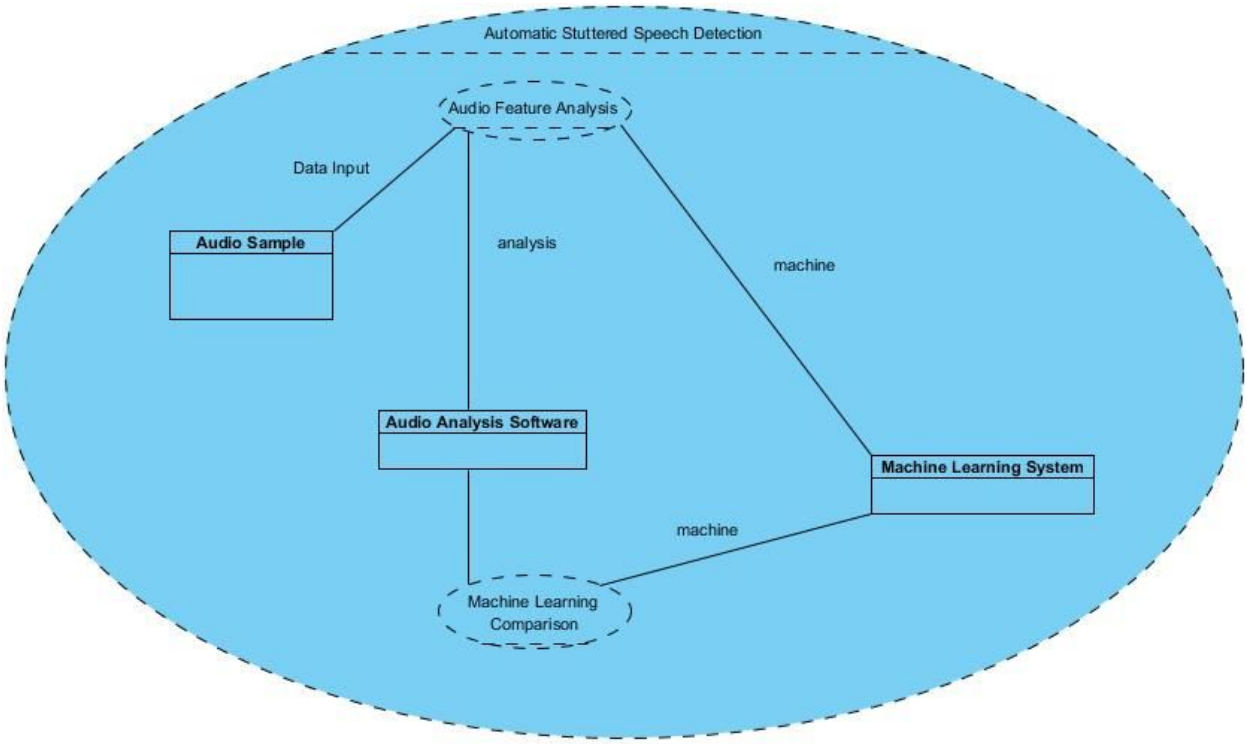
V. ***Appendices***

5.1 Diagrams

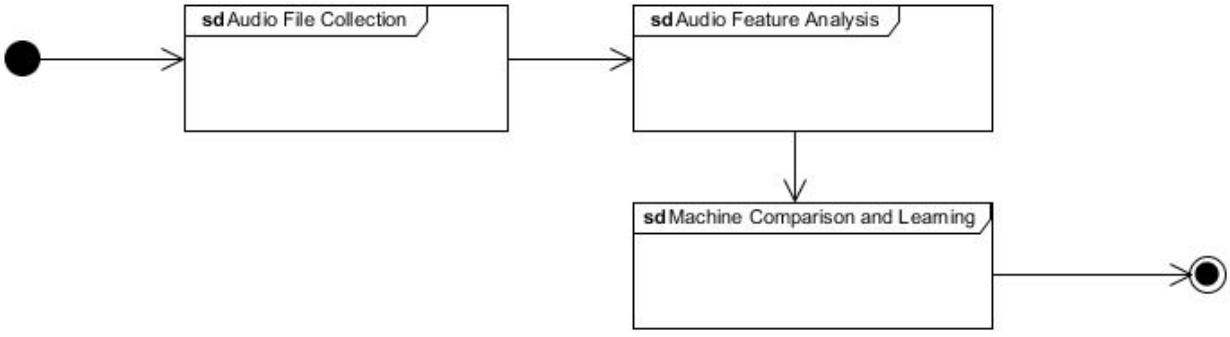
Activity Diagram



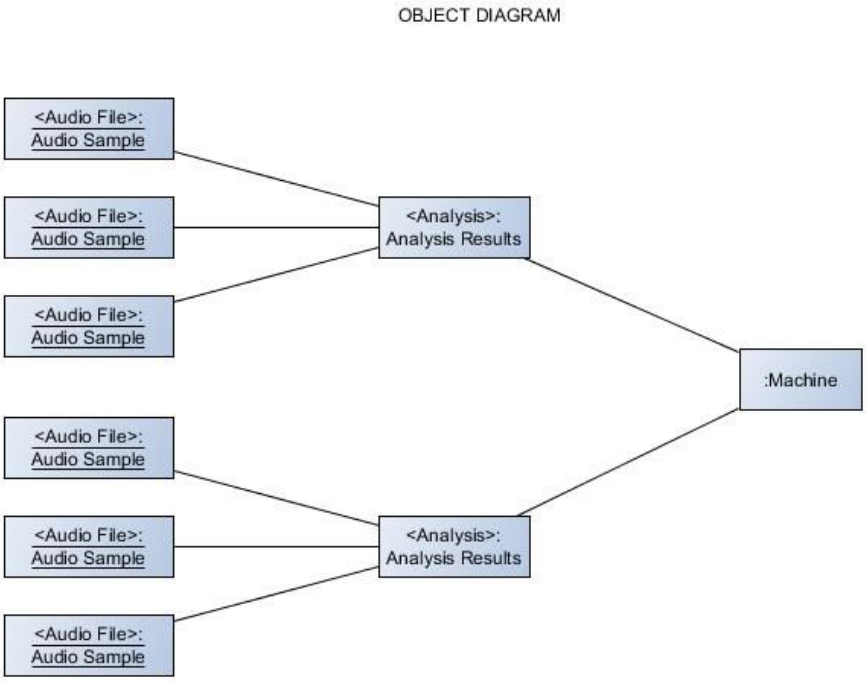
Composite Structure Diagram



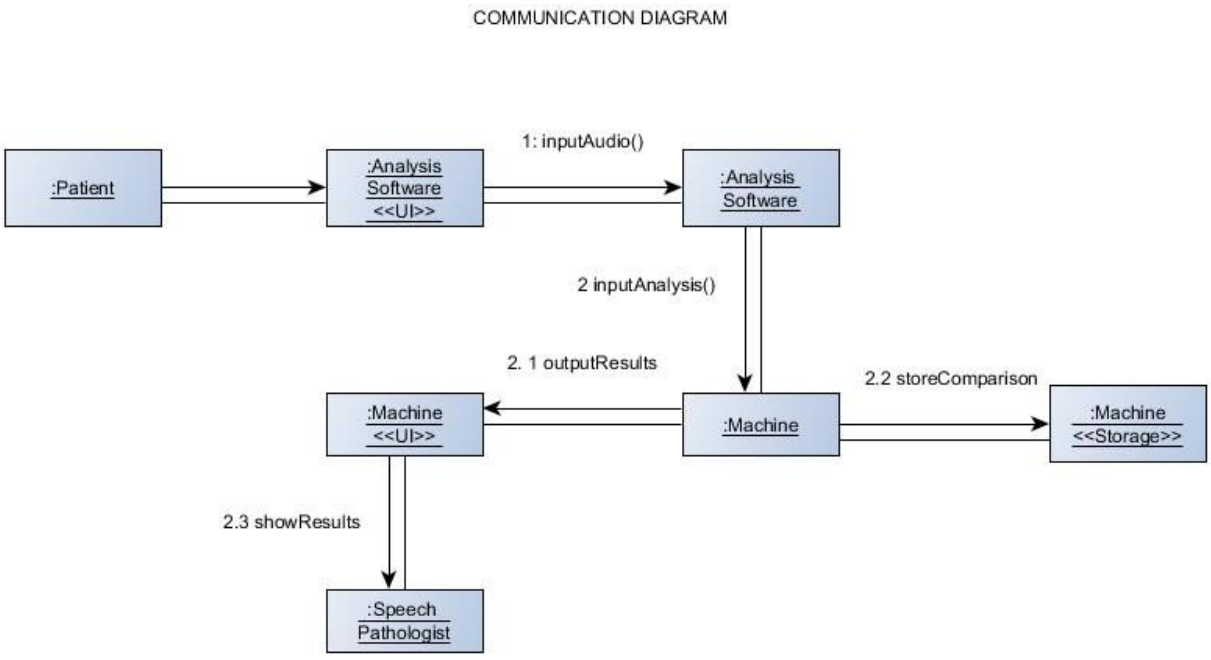
Interaction Overview Diagram



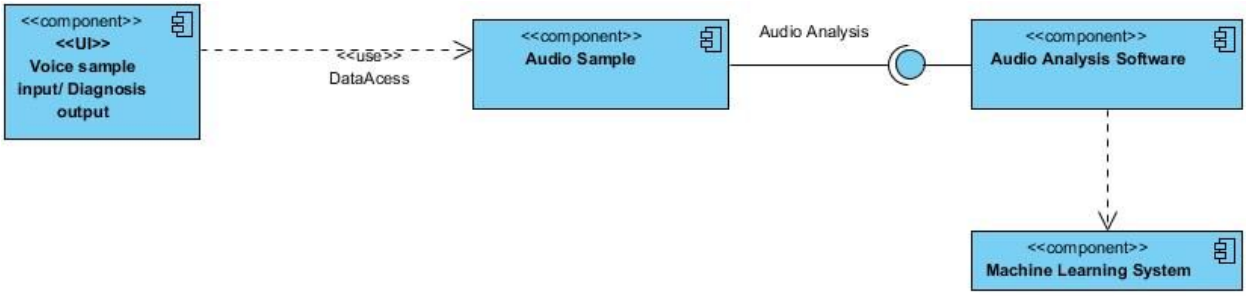
Object Diagram



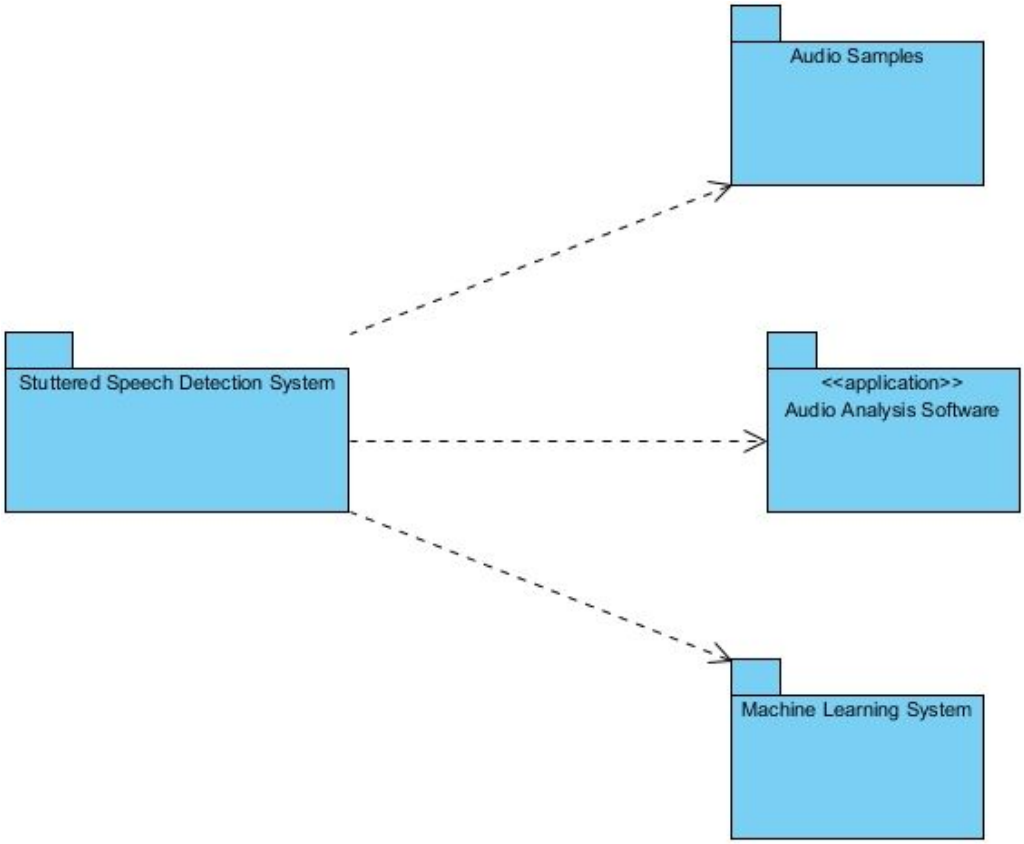
Communication Diagram



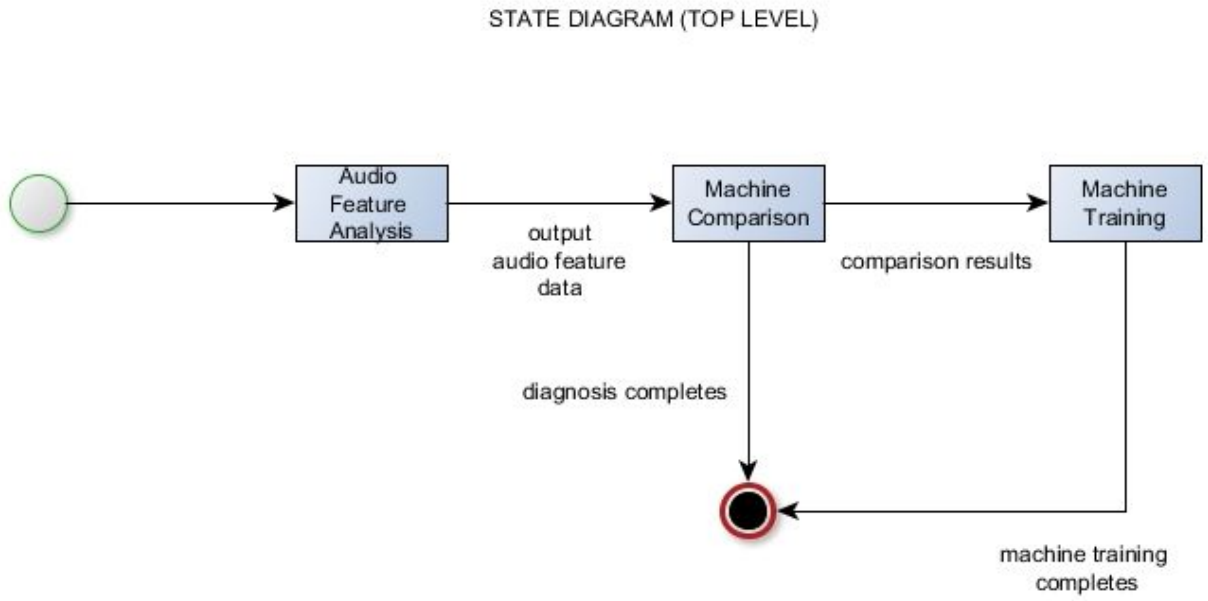
Component Diagram



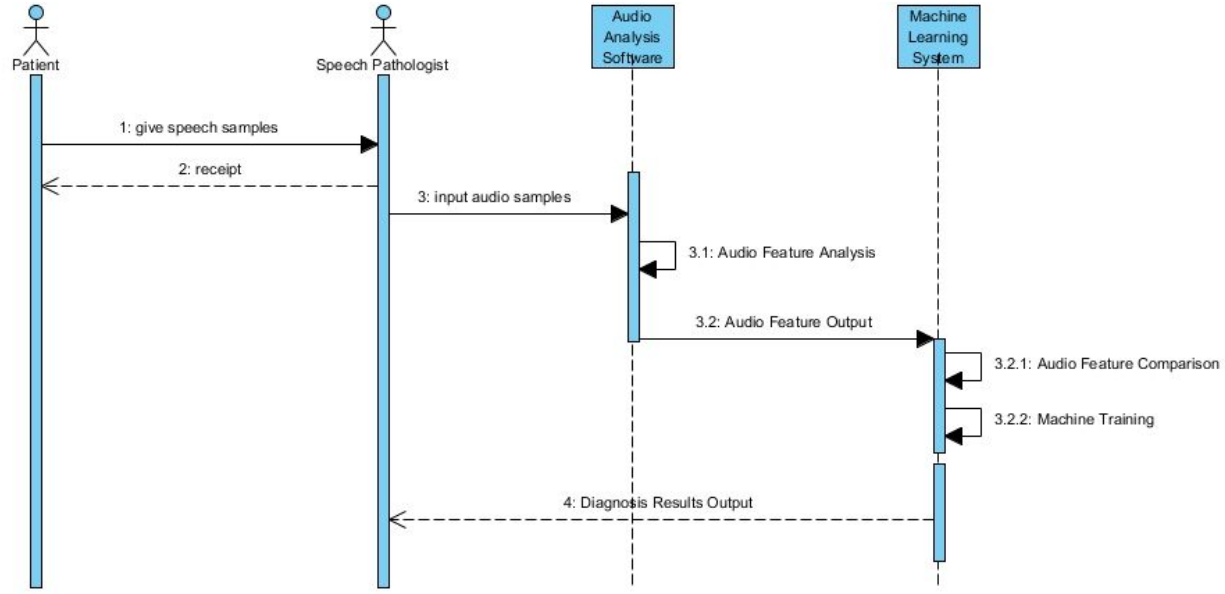
Package Diagram



State Diagram

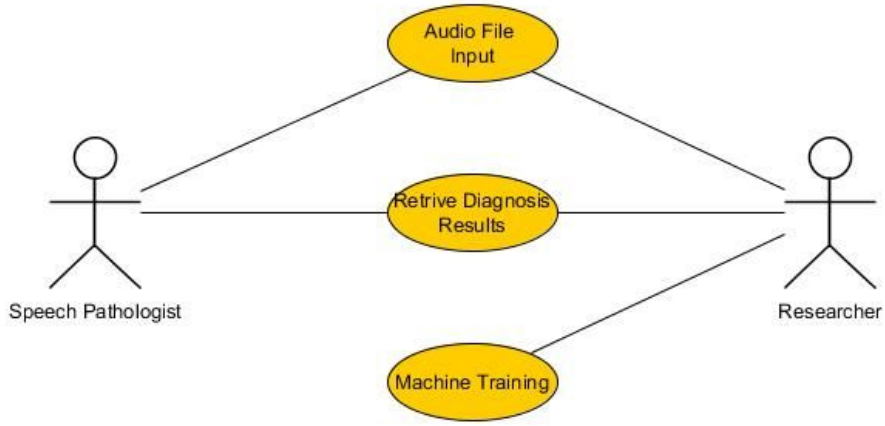


Sequence Diagram

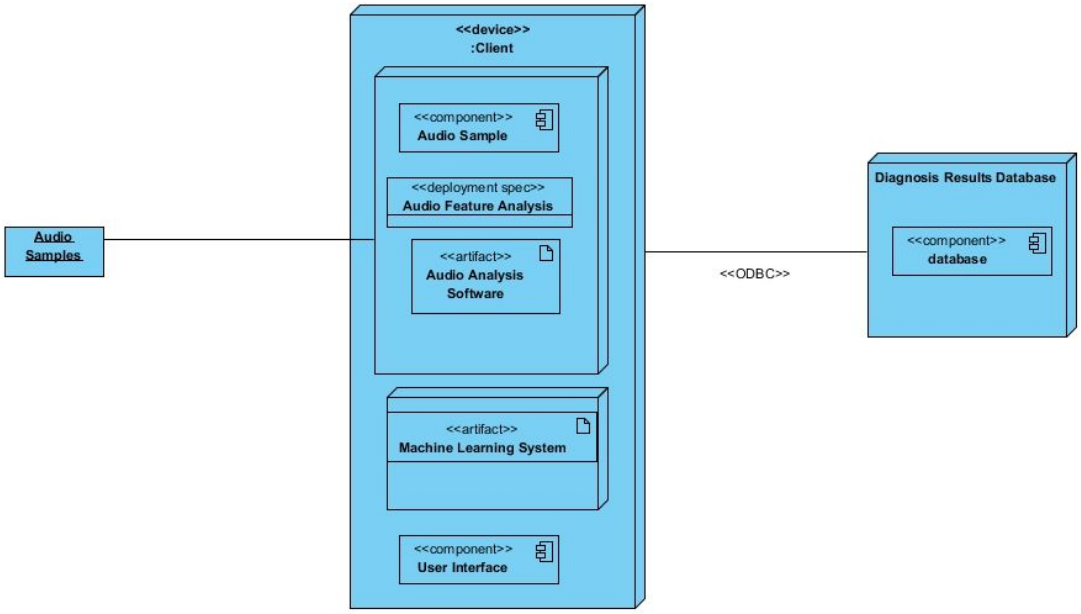


Timing Diagram

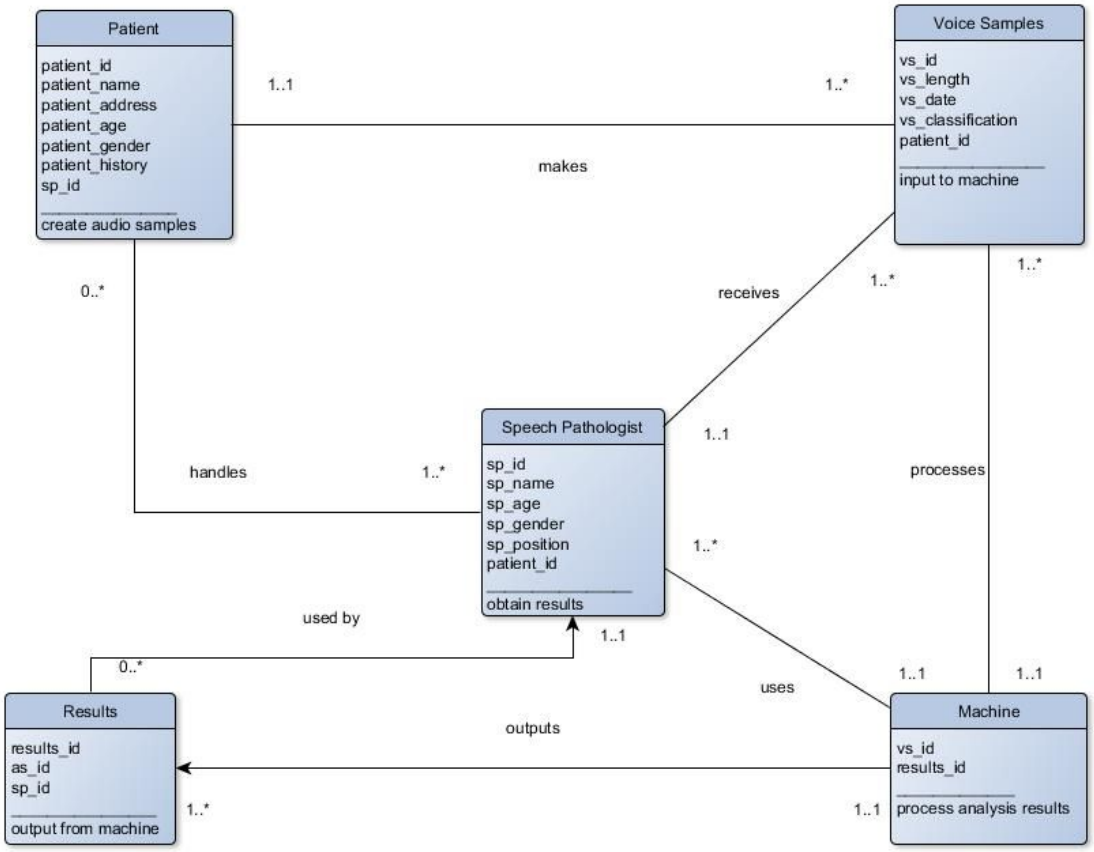
Use Case Diagram



Deployment Diagram



Class Diagram



5.2 Event Table

EVENT	TRIGGER	SOURCE	USE CASE	RESPONSE	DESTINATION
1. Speech Pathologist wants to input voice sample	Input voice	Speech Pathologist	Voice input	Prompt user "Voice Sample received"	Audio Analysis Software
2. Audio feature analysis	Voice input	Audio Analysis Software	Voice Analysis	Prompt user "Audio Feature Analysis being processed"	Machine learning system Speech Pathologist
3. Machine makes comparisons	Audio feature results	Audio Analysis Software	Stuttered speech detection	Prompt "Machine comparison being processed"	Machine Learning System
4. Machine learns	Machine Comparison Results	Machine Learning System	Machine training	Prompt "Machine training being processed"	Machine learning system
5. Speech Pathologist gets results	Machine comparison results	Machine Learning System	Final Results output	Prompt "Voice analysis completed"	Speech Pathologist

Bibliography

Amazon Web Services. *Amazon Machine Learning Documentation*. Retrieved September 5, 2016, from https://aws.amazon.com/documentation/machine-learning/

Arizona Board of Regents. *Fluency Disorders - Communication Sciences and Disorders - Northern Arizona University*. Retrieved September 5, 2016, from https://nau.edu/chhs/csd/clinic/fluency-disorders/

Chee, L. S., Ai, O. C., Hariharan, M., & Yacob, S. (2010). Automatic detection of prolongations and repetitions using LPCC.

Deng, Li; Douglas O'Shaughnessy (2003). *Speech processing: a dynamic and optimization-oriented approach*. Marcel Dekker. pp. 41–48. ISBN 0-8247-4040-8

Geetha, Y. V., Pratibha, K., Ashok, R., & Ravindra, S. K. (2000). Classification of childhood disfluencies using neural networks. *Journal of Fluency Disorders*, 25(2), 99–117.

Hariharan, M., Yaacob, S., Chee, L. S., & Ai, O. C. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, *39*, 2157–2165.

HomeSpeechHome. *Stuttering, everything you need to know in simple terms*. Retrieved September 5, 2016, from http://www.home-speech-home.com/stuttering.html

Howell, P., & Sackin, S. (1995). Automatic recognition of repetitions and prolongations in stuttered speech. *In Proceedings of the first World Congress on fluency disorders.*

Howell, P., Sackin S., & Glenn, K. (1997a). Development of a two-stage procedure for the automatic recognition of disfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical disfluency classifiers. *Journal of Speech, Language, and Hearing Research*, 40(5), 1073.

Howell, P., Sackin, S., & Glenn, K. (1997b). Development of a two-stage procedure for the automatic recognition of disfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers. *Journal of Speech, Language, and Hearing Research*, 40(5), 1085.

Kully, D., & Boerg, E. (1988). An Investigation of Inter-clinic Agreement in the Identification of Fluent and Stuttered Syllables. *Journal of Fluency Disorders*, *13*, 309–318.

Mahesha, P., & Vinod, D. (2012). An Approach for Classification of Disfluent and Fluent Speech using K-NN and SVM. *International Journal of Computer Science, Engineering and Applications (IJCSEA),* 2(6), 23–32.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Nanopoulos, A., Alcock, R., & Manolopoulos, Y. (2001). Feature-based Classification of Time-series Data.

Practical Cryptography. (2009). *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. Retrieved September 5, 2016, from

http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Ravikumar, K., Reddy, B., Rajagopal, R., & Nagaraj, H. (2008). Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, *2*(10), 2142–2145.

Szczurowska, I., Kuniszyk-Jozkowiak, W., & Smolka, E. (2006). The Application of Kohonen and Multilayer Perceptron Networks in the Speech Nonfluency Analysis. *Archives of Acoustics*, *31*(4), 205–210.

YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, B.Mathieu, S.Essid, T.Fillon, J.Prado, G.Richard, proceedings of the 11th ISMIR conference, Utrecht, Netherlands, 2010.