TheraSpeech: An Intelligent Speech Therapy System for Velopharyngeal Insufficiency

Joanne C. Castillo, Joseph Z. Fabian, Mark Lean V. Magbanua, John Rhio F. Saloma, and Ria Sagum

***Abstract*—This document outlines the use of an intelligent speech therapy system as a learning aide for individual having a speech disorder known as Velopharyngeal Insufficiency. The goal is to detect speech disorder under a phoneme specific etiology such as plosives, affricates, and fricatives. The system used speech recognition to capture and evaluate speech signals and give an accurate and correct module for the user. The digital signals were processed, analyzed and computed using Mel Frequency Cepstral Coefficient (MFCC). Through these methods, system will come up with a speech evaluation results and modules for therapy. The system was tested by random students having an accuracy of 85.55%.**

***Index Terms*—MFCC, velopharyngeal insufficiency, phonetically-balanced words, silence removal, pre-emphasis, framing, humming, windowing, fast fourier transform (FFT), mel filter bank.**

I. Introduction

TheraSpeech is an intelligent system for a speech disorder known as Velopharyngeal Insufficiency (VPI)[1]. Since speech assessment is a complex process, the method of assessing, describing, and interpreting an individual's communication ability requires the integration of a variety of information gathered in the evaluation process [1]. The main goal of the system is to extract the speech signal using the MFCC as the transformation domain to provide an accurate speech assessment evaluation and module for therapy.

One of the most common cause of VPI is cleft palate, other than that, structural problem, neurogenic cases, mechanical interference and phoneme-specific difficulties also contributes to this kind of disorder. Focusing on the phoneme specific etiology of the disorder, different manners of articulation such as plosives, affricates and fricatives must be considered [2].

Nowadays, where technology is one of the key to advancement, speech recognition became one of the subsets in the artificial intelligence field which translates gathered data into meaningful information and learns from the knowledge representation to communicate to the user.

Considering these factors and problems led to the way on developing an intelligent therapy system that will asses the oral communication ability of an individual and serves as a learning aide for those who have VPI. Since the study focuses on the phoneme specific etiology, existing foreign and local standardized therapy software are used as a guide for the assessment and evaluation of speech for the system.

Although several therapy systems have been completely developed over times, these systems were specifically designed only for the portion of assessing the communication abilities of the user. Thus it covers only the initial part of the therapy for VPI and there are no limited studies found regarding any system that provides the next process after the assessment phase.

II. Related Works

Determining the kind of articulation error will help the pathologist on what treatment they will provide to the person suffering in VPI. More often, pathologist gives diagnostic articulation inventories which are not really tests, but sets of pictures, words, nonsense syllables, or reading materials that can be used to elicit speech samples to be analyzed for error [3].

TABLE I: Percentage of Consonants Correct (PCC)[2]

Severity	Percentage
Mild	>85%
Mild-moderate	67-84%
Moderate-severe	50-66%
Severe	< 50%

The percentage of consonants correct (PCC) is derived by dividing the total number of correct consonants by the total number of consonants in the targeted words. PCC values correspond to an ordinal severity scale that is useful for describing intelligibility and impairment. PCC values can be assigned four levels of severity of involvement using the severity adjectives shown in Table I [4]. It is used to determine whether the user needs to undergo speech enhancement.

Computer based speech training system helps people in their caseloads improve their communication skills, and the visual feedback helps them successfully develop and remediate speech errors. It proves that intelligent system plays an important role in speech therapy [5]. An interaction with therapy software is also more dynamic and responsive

Manuscript received February 15, 2014; revised July 30, 2014.

J. C. Castillo, J. Z. Fabian, M. L. Magbanua, and J. R. Saloma are with the Department of Computer Science, Polytechnic University of the Philippines in Sta. Mesa, Manila 1008 Philippines (e-mail: castillojoanne0208@yahoo.com; joseph_fabian@yahoo.com; markmagbanua@rocketmail.com; jrhio08@yahoo.com).

R. A. Sagum is with the Department of Computer Science, College of Computer and Information Sciences, Polytechnic University of the Philippines in Sta. Mesa, Manila and is also with the Faculty of Engineering, University of Santo Tomas in Manila, Philippines (e-mail: rasagum@pup.edu.ph).

[1] VPI stands for Velopharyngeal Insufficiency that occurs when there is an abnormal coupling of oral and nasal cavities during speech especially with vowel productions due to structural, neurogenic and/or behavioral tissues

[2] PCC= Consonants Correct / Total Consonants In Target Words.

than the paper-based exercises that therapists often set for the patient to complete. Computer-based therapy systems also provide a rich source of data for speech and language therapy research purposes, including data that human therapists cannot gather, such as the patient's reaction times [6].

III. TheraSpeech Resources

A. Phonetically-Balanced Words

The proponents utilized a word database called Assessment Module and Therapy Module. This database which contains most occurring phonetically-balanced words of year 2012 were utilized and identified by pattern matching. The module varies from the age calculated by the system. Each module contains words as well as their description which were used to evaluate the patient's severity of disorder. Words are categorized through different priorities covered by phonemes: plosives, affricates and fricatives. Table II shows a sample of word database for Module 1.

TABLE II: Sample Word Database of Module 1

Word	Description	Phoneme	Priority
Apple	A red juicy and tasty fruit	P	Medial
Basket	A container made of dried coconut leaves or plastic.	B, K	Initial, Medial
Bread	A delicious pastry.	D	Final
Word	Description	Phoneme	Priority

B. Photo-Articulation Tool

The system utilizes photos that would represent the words in the assessment phase. Plain white backgrounds, clear and easy to identify photos were used to as the visual tool.

C. Rules

The system follows the arbitrary methods in assessing and treating user's utterance. This method comes in different level such as direct identification, description, with choices and direct imitation. Errors in utterance are classified through the provided levels providing the summary result of mispronounced phonemes.

IV. Speech Signal Processing Using MFCC

In this work, the proponents utilized two speech recognition algorithms to identify and analyze speech signals. The input speech waveforms are processed using the built-in Speech Recognition Engine (SRE) and the Mel Frequency Cepstral Coefficient (MFCC) for the transformation domain. Fig. 1 shows the system architecture of the TheraSpeech in the speech recognition process.

The system will start on assessing the user's speech through Photo Articulation Test from the Photo Articulation Database. Once the user is evaluated and advice to take the therapy, the result from the assessment will be used for accessing the module for therapy. String matching will be the algorithm for getting the appropriate module for the patient.

Taking consideration on the recognition performance of the SRE wherein it has a higher chance of getting a correct recognition response even if the words are not pronounced properly. Some instance is when the system asked the user to pronounce the word Butterfly. Using the built-in SRE only, the words utterly, terfly or fly can be considered correct. But since we are intended of getting an accurate recognition to assess the speech and to treat VPI of certain individual, we used MFCC which uses different processes and algorithms that filters unnecessary elements to come up with a more concise spectrum signal of the input waveforms.

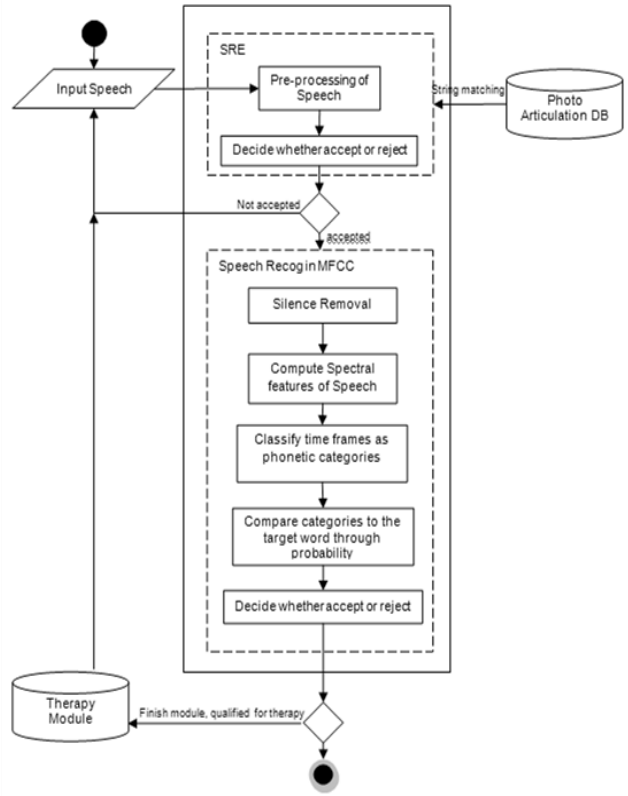

Fig. 1. System architecture.

Before the two recognition algorithm takes place, the system utilized Silence Removal to eliminate the stillness on the speech signal and moved forward all the detected waveforms with a higher spectrum. The Fig. 2 and Fig. 3 below show how the Silence Removal took place on the speech signal.

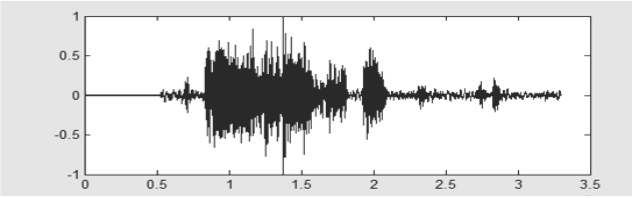

Fig. 2. Original speech signal of the word *Grasshopper.*

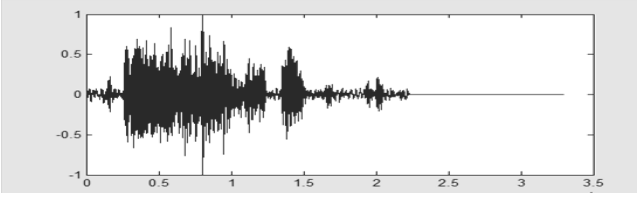

Fig. 3. The word *Grasshopper* after removing silence.

The main aim of removing the silence is to prepare the speech signal for the two recognition algorithm afterwards. The process in removing silence is to divide the spectrum

from each frames ranging to 25ms. After this, the maximum amplitude on each frame must be analyzed. If the amplitude on each segmented frames doesn't exceed at 0.5ms, the frame will be transferred at the end of the speech signal.

After silence has been removed from the speech waveform and been simplified, this speech waveform are converted into digital signal form to produce digital data representing each level of signal at every discrete time step. The digitized speech samples are then processed using MFCC to produce speech features. This MFCC which is the transformation domain used for the computation of the cepstral coefficient from the speech signal will be next process to employ.

This stage is the most important in the entire process since it is responsible for extracting relevant information from the speech frames, as feature parameters or vectors to produce a better recognition performance.

MFCC consists of seven computational steps. Fig. 4 shows the process of speech signal through MFCC [7].

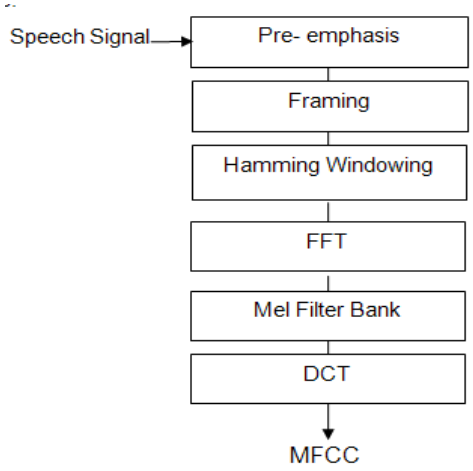

Fig. 4. Speech signal extraction using MFCC.

After extracting the signal, the next step is to calculate the mean squared error between two signals using their MFCC values to be followed by the system's decision.

V. Performance of TheraSpeech: Results and Analysis

The system was initially tested by the selected students coming from different sections of Kindergarten to Grade-4 in Tandang Sora Elementary School-Quezon City S.Y. 2013-2014. Also, the system was tested and rated by the experts in the field of Language and Speech Pathology in terms of: Accuracy, User-Friendliness, and Responsiveness Content.

Each module contains atleast 40 words for the Assessment Phase depending on the age of the speaker. A total of 943 words are tested during the implementation of the system.

The test showed that the system has an 85.55% of accuracy rate. The formula below shows the formula used for the accuracy. Number of correct assessment refers to the cohesion between manual assessment and machine assessment. Number of words refers to the total number of words provided in all assessment.

$$\text{Accuracy} = \frac{\sum \text{no. of correct assessment}}{\sum \text{number of words}} \times 100\%$$

VI. Conclusions and Recommendations

Overall, based on the findings of the study, the researchers' conclusions were drawn.

First, the system is responsive and user-friendly based on the gathered data among the respondents. Also, it is proven that the content of the system is effective for the users of the system.

Second, there were problems of the software while in the implementation phase that needs to be fixed. The MFCC needs to be improved to achieve higher recognition rate. And also, when using a build in algorithm, make sure that you have a deeper knowledge about the tool to avoid low complexity of the system. Patching algorithm with another algorithm may cause flaws if not utilized effectively. In this study, the system acquired an 85.55% of accuracy which maybe greater if merge another algorithm that will work on the speed of comparing the signals in the system.

These are the following suggestions and recommendations which might help the future researchers to develop the existing system:

1) Add some features of the system such as recording the assessment phase which is useful for the speech pathologist when reviewing the utterance of the patient.
2) Include other letters and combination of letters such as clusters, diphthongs, glide and nasals.
3) Improve the accuracy of the speech recognition for better results.
4) Improve the size of the database to allow pathologist add new words or module in the system.

Acknowledgement

The proponents want to express their sincere and abundant gratitude to the people who made this research possible.

First and foremost, we would like to thank God that right from the beginning of our work, He helped us and filled us with His everlasting goodness and guidance.

We would like thank Prof. Ria Sagum, our Thesis adviser for the assistance, guidance and encouragement to pursue this study.

We also like to acknowledged, Mrs. Joanne Rabang, speech pathologist of Chatter Therapy Clinic, who guided us to improve the content and functionality of the program. Thank your for giving an ample time with us despite the fact that you are having a hectic schedules with your patients.

To our beloved friends and classmates, BSCS 4-4, thank you for providing the resources that we need during the study and believing that we can finish it.

Lastly, to our ever supportive family, thank you for all the understanding, assistance, financial support and advice.

References

[1] ASHA. Directory of Speech-Language Pathology. Assessment Instruments Introduction. (2009). [Online]. Available: http://www.asha.org/SLP/assessment/Assessment-Introduction/
[2] M. K. Berger *et al*., "Instrumental assessment of velopharyngeal dysfunction: multi view video fluoroscopy vs. nasopharyngoscopy," University of Michigan Health Systems, C.S. Mott Children's Hospital, 2011.
[3] V. C. Riper, "Speech correction: an introduction to speech pathology and audiology," National Institute for Mentally Handicapped, 1996.
[4] *Speech Sound Assessment and Intervention Module*.
[5] B. Grawemeyer, R. Cox, and C. Lum, "AUDIX: A Knowledge-based System for speech- therapeutic auditory discrimination exercises," in *Proc. MIE2000 and GMDS2000*, 2000, vol. 77, pp. 568-572.

[6] Video Voice. [Online]. Available: http://www.videovoice.com/vv_benies.htm
[7] H. Holmes, *Mel Frequency Cepstral Coefficient*, 2001.

Joseph Z. Fabian was born on August 6, 1993 at Quezon City. He finishes his secondary education in Ismael Mathay Sr. High School. He took up bachelor of science in computer science in Polytechnic University of the Philippines, Sta. Mesa, Manila.

He practiced his skills as a database administrator in Nissan Gallery, Quezon Avenue during their on-the-job training.

Mr. Fabian ranked as the third place in the CCMIT-PUP Movie Making Contest. He also became a participant of different school seminars including Ideaspace Technopreneur Competition, PUP Download Feast, 3rd CCIS-DCS Computer Science Research Symposium and 8th Bi-Annual TechDoc Convention. Mr. Fabian is also a feature writer of the compiler- a computer science official school paper.

Joanne C. Castillo was born on February 8, 1993 at Commonwealth. She graduated from Toro Hills Elementary School and finished her secondary education in Ismael Mathay Sr. High School. She took up bachelor of science in computer science in Polytechnic University of the Philippines in Sta. Mesa, Manila.

She worked as a database administrator in Nissan Gallery, Quezon Avenue during their on-the-job training.

Ms. Castillo is a news writer of their college publication, The Compiler. She attended seminars such as 8th Bi-Annual TechDoc Convention, Ideaspace Technopreneur Competition, PUP Download Feast and 3rd CCIS-DCS Computer Science Research Symposium. She receive an award as a first place in the CCMIT-PUP Movie Making Contest.

John Rhio F. Saloma graduated at Palatiw Elementary School and finished his secondary education in La Immaculada Concepcion School in Pasig City. He took up bachelor of science in computer science in Polytechnic University of the Philippines, Sta. Mesa, Manila.

He practiced his skills as a system developer in UP-DILC (Diliman Interactive Learning Center) during their on-the-job training (OJT).

Mr. Saloma became a participant of different school seminars including PUP Download Feast, 3rd CCIS-DCS Computer Science Research Symposium and 8th Bi-Annual Technical Documentation Convention in Baguio.

Mark Lean V. Magbanua finished his primary and secondary education in La Charity School of Antipolo. He took up bachelor of science in computer science in Polytechnic University of the Philippines, Sta. Mesa, Manila.

He worked and practiced his skills as a system developer in UP- DILC (Diliman Interactive Learning Center) during their on-the-job training (OJT).

Mr. Magbanua became a participant of different school seminars including PUP Download Feast, 3rd CCIS-DCS Computer Science Research Symposium and 8th Bi-Annual Technical Documentation Convention in Baguio.

Ria A. Sagum was born in on August 31, 1969 at Laguna, Philippines. She took up bachelor in computer data processing management from the Polytechnic University of the Philippines and Professional Education in Eulogio Amang Rodriguez Institute of Science and Technology. She received her master degree of computer science, in De La Salle University in 2012.

She is currently teaching at the Department of Computer Science, College of Computer Management and Information Technology, in the Polytechnic University of the Philippines in Sta. Mesa, Manila and a lecturer at the Information and Computer Studies, Faculty of Engineering, in the University of Santo Tomas in Manila.