Analysis on Using Synthesized Singing Techniques in Assistive Interfaces for Visually Impaired to Study Music

Kavindu Ranasinghe, Lakshman Jayaratne

Abstract—**Tactile and auditory senses are the basic types of methods that visually impaired people sense the world. Their interaction with assistive technologies also focuses mainly on tactile and auditory interfaces. This research paper discuss about the validity of using most appropriate singing synthesizing techniques as a mediator in assistive technologies specifically built to address their music learning needs engaged with music scores and lyrics. Music scores with notations and lyrics are considered as the main mediators in musical communication channel which lies between a composer and a performer. Visually impaired music lovers have less opportunity to access this main mediator since most of them are in visual format. If we consider a music score, the vocal performer's melody is married to all the pleasant sound producible in the form of singing. Singing best fits for a format in temporal domain compared to a tactile format in spatial domain. Therefore, conversion of existing visual format to a singing output will be the most appropriate nonlossy transition as proved by the initial research on adaptive music score trainer for visually impaired [1]. In order to extend the paths of this initial research, this study seek on existing singing synthesizing techniques and researches on auditory interfaces.**

Keywords-singing synthesizing; auditory interfaces; assistive technologies

I. INTRODUCTION

This Rapid advancement in science and technology has made humans day to day life easier. But not all the people get the benefits equally. The research focuses on assisting a special group of differently abled people. Almost all of the visually impaired population tends to choose music as their aesthetic subject in their primary and secondary education. Visually impaired people are hearing specialists and are brilliant music performers. But, music notation scripts, the main mediator in musical communication channel between a composer and a performer is not in a friendly format for them. Visually impaired people require third party assistance to convert visual music notations to a format readable for them. Even when printed forms of music notation scripts are converted to music Braille, they find it difficult to follow. In earlier steps of learning curve, more priority is given to lyrics. Even music teachers tend to teach lyrics singing first and note singing after that, because meaning of lyrics builds a logical combination which is easy to memorize, compared to a sequence of music notes. The evaluation results initial research carried out in defining an adaptive score trainer for visually impaired, also depicted that singing output proposed by that initial research solution is the best output method compared to existing music Braille output with 81% user votes.

Singing best fits for a format in temporal domain compared to a visual format such as a notation and lyric script in spatial domain. But the expected outcome of the research is not just a synthesized singing, but an interface which is capable to successfully communicate all forms of information that can be embedded within a music notation script through auditory objects. Some of the recent researches on Auditory Displays which examine how human auditory system can be used as the primary interface channel for communicating and transmitting information also reveal its massive potential in improving assistive technologies.

The main qualities that a good singing synthesizing technique target are naturalness, intelligibility, tonality and rhythm. Naturalness and intelligibility is common for both speech synthesis and singing synthesis while tonality and rhythm is specific for the latter. Singing synthesizing techniques which has been already explored throughout the research history can be included under the main categories concatenation synthesis, formant synthesis, articulatory synthesis, Hidden Markov Model (HMM) based synthesis and sine-wave synthesis. Concatenation synthesis is derived by aggregating units of sound signals such as words, phrases, syllables, phones, di-phones, and half phones. Formant synthesis use acoustic models to re-build the shape of signals. Models of the human vocal tract and the articulation processes are used in articulatory synthesis. In HMM-based synthesis effect of frequency spectrum, fundamental frequency and duration are modeled simultaneously while sine-wave synthesis use pure tone whistles and have less chance in re-creating naturalness.

In existing literature, we come across many types of singing synthesis techniques. Most famous singing synthesizer VOCALOID is built upon concatenation based synthesis. VOCALISTENER is a tool based on formant synthesis and SPASM is an example for articulatory synthesis. LYRICOS uses a sinusoidal model based on sinewave synthesis while Sinsy uses HMM-based techniques. Especially the researches by MIT and KTH have contributed a lot to the knowledge related to singing synthesis. But synthesized singing has not yet been focused as a mediator in musical communication for visually impaired and also the effective use of signal

DOI: 10.5176/2251-3043_4.4.342

processing capabilities on interfaces for visually impaired is still at research stage.

The outcome of this study helps to fill the existing accessibility gap which visually impaired music lovers face and is expected to enhance the opportunities they have to study and experiment on music independently.

II. Related Work

A. Singing Synthesizing Techniques

Singing synthesizing is not a new topic for researches. It has been evolved from 1950s. In earlier stages of its evolution cycle, most inventions were based on hardware which mimics the behavior of vocal tract. The physical models may range from simplest structures such as acoustic tube to advanced solutions such as VOCODERS. Physical models in earlier ages of evaluation were expensive and were not affordable. Even though the knowledge has transferred from hand to hand for years, still the naturalness of the latest and most advanced solution is questionable.

Major types of singing synthesizing approaches can be categorized as physical models and spectral models. But our focus in this paper mostly runs upon the spectral models. Spectral models can be further categorized based on the type of inputs and outputs accepted in each approach. The first type accepts score and lyrics as input to generate a synthesized singing. Second type accepts speaking signal as input and coverts the same to singing. It is referred as speech-to-singing (S2S) synthesis in literature. Last type accepts singing as input and re-generates a singing output again with some modifications executed upon the signal.

Most challenging task in singing synthesis is replicating the natural intonation pattern. In most of the approaches the naturalness is bit destroyed due to the inability in re-synthesizing prosodic variables as it is. Some phenomenon originally built for different purposes in different disciplines such as Linear Predictive Coding (LPC), Dynamic Time Warping (DTW) are also used in singing synthesizing practices.

i) Concatenation Synthesis

From the early Vocoders developed in bell labs, human voice synthesizers have developed up to singing synthesizers and have become a trendy research area in last few decades. VOCALOID[2] is a singer synthesizer application developed by Yamaha. According to creators of the tool, the singer library has been created using recorded combinations of vowel and consonant sounds, and many combinations of different pronunciations (variations in nasal quality, etc.) and song lyrics sung by live singer in a studio. The recorded data is broken into sound fragments and further adjusted, edited, and refined into elements suitable for concatenation into a smooth sound. These elements are then stored to a database known as a singer library or singer database. The synthesis engine in VOCALOID concatenates the voice elements to generate the singing voice. The engine receives MIDI messages from the score editor; extracts the score data, lyrics, expression data, and other necessary information from these messages; retrieves the necessary sound elements from the singer library; and concatenates these elements to create the singing.

CANTOR 2 is a real-time vocal synthesizer created by VirSyn Software Synthesizer. VOSE (Vocal Simulation Engine) is the vocal synthesis technology used in CANTOR. When the lyrics (in plain English) and melody is given as inputs corresponding singing can be synthesized using this tool. Simulation Engine uses a combination of additive synthesis and noise sculpting in generating sound. It offers in built voices and create user defined individual voice characters.

The Whistler (Windows Highly Intelligent STochastic taLkER) Music Synthesizer[3], under Microsoft research is another most successive singing synthesizer that we can found including both male and female voices. Whistler uses a simplified database of about 3,000 allophones, which were isolated by cutting digital waveform recordings of the human voice into sections.

One of the major problems in merging signal chunks is addressed in [4]. A large database of acquired vocal sounds was used to produce more complex phoneme-based singing synthesis free with glitches. Applications described in the paper exposes a potential to create rich reservoirs of articulations and timbre identities having the aim to tap into greater timbral variety for the sonification work. The improvements to Frequency Modulation (FM) vocal synthesis detailed in this paper can be extended to other audio rate modulation schemes, especially those which employ single modulator / multiple carrier structures.

In deep experimenting the singing synthesizing field; the study [5], is discussing about proposing a score to singing synthesis system and as the method, score will be written in a score editor and saved in the MIDI format. As the study suggests Singing Voice Synthesis (SVS) become a trendy last few decades and it discusses about the urgency and significance of using SVS systems while describing the characteristics of a SVS system. At the same time it contrasts and compares speech and singing processes as well as the vocal music generation and instrumental music generation. The study discussed about the researches done to improve the natural quality of the sounds generated.

Since the system [5] is generating singing synthesis system researchers have discussed more about generating an audio out put through MIDI file. Since the study focuses basically on providing a better way to produce synthesized singing, validity of using software like MBROLA and MIDI file formats and converting methods will be important to consider about. [5] focuses more on lyrics singing with intelligibility than simulating the sense of naturalness. The use of concatenate synthesis in singing synthesize is discussed in [5] and the previous project failures happened in the past in synthesized singing and the reasons for failures. As an example Swedish use of MBROLA output was not much natural because their diphone database was derived from spoken language. According to the study; most of the diphone databases are available in male voice. That itself suggests as a drawback of using available databases.

In deep surfing score to audio conversion there are considerable amount of researches that being carried out with dynamic time warping to score to audio alignment. As both [6] and [7] suggests it is one of the best ways to minimize the error rate that can be occurred in converting score to audio. The research [7] presents a novel algorithm to align two sequences by time warping them optimally minimizing the weaknesses of Dynamic time warping those are cost and memory limitations using iterative approach. The evaluation is done between a score in MIDI format and instrumental performance of the score as the audio. The algorithm "Short-Time Dynamic Time Warping" (STDTW) in study [7] had evaluated in aligning musical score and the time axis of audio performance of it. So the aligning results may be important for us in measuring the accuracy of the conversion music score to audio. As the problems in the study they have are, impossibility of identifying the exact tempo which is not constant and the notes played may be different than written in the score and in polyphonic music it says difficult to understand the note playing. In monophonic compositions there is only one frequency allocated for a specific point of time. But in polyphonic music compositions, there can be more frequencies played parallel making it more difficult in identifying the exact tempo. The core of the study is highly focused on mathematical algorithms.

ii) Formant Synthesis

A novel parametric model of spectral envelope to produce realistic variations of intensity in recorded or synthetic vowels has been presented in [8]. Researchers specifically propose the use of 4-poles resonators to synthesize the vocal formants, instead of 2-poles resonators. Using this parametric model, the researchers were able to define a set of functions describing the variation of parameters along intensity in singing voice.

[9] Proposed a new method of automatic vocal harmony that is significant because, unlike existing methods, it is suitable for singers without a good sense of rhythm yet does not sacrifice the quality of consonance. The test results reveal that this approach is successful in achieving a better level of perceived harmonic consonance, transitional smoothness, as well as overall naturalness and pleasantness.

An approach using a singing performance to control the expression of a Singing Voice Synthesizer is introduced in [10]. It aims at improving the process of generating expressive synthetic singing, for which most state-of-the-art systems rely on modeling techniques or direct manipulation of low-level parameters. Researchers have introduced the performance analysis module of their proposed system that extracts pitch, dynamics and vibrato information from an input voice signal. Given lyrics and a voice signal, the system can automatically perform phonetic segmentation from which the phonetic timing of the synthesis can be derived. Due to sound quality limitations of the SVS used, they have suggested two alternatives for the phonetic alignment besides a one-to-one alignment with the input voice. First experiments showed that best sounding quality and natural sounding singing is achieved aligning voiced phonemes only. Although their final aim was to implement a real-time control of SVS, the current off-line implementation already offers a direct and easy way of controlling synthetic singing.

A Four tone synthesizing model has been brought into discussion in this study [11]. Four-tone is one of the characteristics of Chinese language. The pitch fluctuation pattern is classified into four main types in Chinese. The pitch fluctuation pattern of 1st tone is flat high pitch. That of 2nd tone is a tone starting with mid pitch and rising to a high pitch. That of 3rd tone is a low tone which dips briefly before raising a high pitch. That of 4th tone is a sharply falling tone starting high and falling to a low pitch.

[12] reconsidered significant F0 fluctuations in singing-voices perception as characteristics that need to control the F0 contour of singing-voices and develops an F0 control model that can control F0 fluctuations. Main conclusions of the research depicts that the naturalness of singing-voices decreases when removing each F0 fluctuation from F0 contours, the naturalness of synthesized singing voices increases by adding each F0 fluctuation into the Melody component and the quality of synthesized singing voices is almost the same as that of real singing voices. The results of the research also show that F0 fluctuations, especially overshoot, vibrato, fine-fluctuation, and preparation are important factors in singing-voices perception. It is capable to generate F0 contours including all F0 fluctuations by determining optimal parameters for each F0 fluctuation.

iii) Formant Synthesis (speech-to-singing)

[13] Proposes a speech-to-singing synthesis system that can convert speaking voices to singing voices by adding acoustic features of singing voices to the F0 contour and spectral envelope and can lengthen the duration of each phoneme. The experimental results shows that it is capable to synthesize singing voices whose naturalness is close to that of actual singing voices and that the F0 fluctuations are more dominant acoustic cues than the spectral characteristics in the perception of singing voices. Although the system has been built for conversion from speaking voices to singing voices, the tests has proved that the same system can be used for conversion from singing voices to singing voices. The researchers points out the applications of the latter in changing the characteristics of singing voices that have already been recorded. Researchers also have plans to use it as a singing-voice effects processor for music productions since can can independently control the F0 levels and spectral envelope. According to the authors, future work will also include research on investigating acoustic features that affect perceptions of the singer's individuality and singing styles and extending system to express them.

iv) Formant Synthesis (singing-to-singing)

[14] Presented a singing voice conversion method based on many-to-many Eigen Voice Conversion (EVC) and training data generation using a singing-to-singing synthesis system. Method brought forward in this research is capable of converting the singing voice quality of an arbitrary source singer into that of an arbitrary target singer by adapting a small number of adaptive parameters of a conversion model using an

extremely small amount of source and target singing voice data. This method also can alleviate the burden of having to record singing voices to develop parallel data sets.

v) HMM-Based Synthesis

In converting visual script in to speech/singing synthesis, the study [15] attempts to develop a framework which analyze and synthesize singing voice. In order to accomplish the goal it estimates source filter voice model parameters by considering both physical and expressive factors and models the dynamic behavior of these features over time using a Hidden Markov Model. Then while in progress of the research it describes the framework using mathematical algorithms and discusses the features of the framework. The other application the [15] suggests for the framework is low-bitrate singing voice coding. And it claims that the previous researches which the study was built upon the compression advantage of encoding vowels using static templates. Hidden Markov Model is defined as "A statistical-tool used for modeling generative sequences characterized by a set of observable sequences." [16]

The research study [17] is attempting to give a solution for the problem of reading a Chinese music score type which does not present the duration of each note in the music script (rhythmic immeasurability problem) by automatically interpreting it in staff. In order to achieve this they have first concluded rhythmic patterns and then using HMM and mathematical calculations have been done.

[18] Study specifically analyzes and model vibrato expressing in order to synthesize Mandarin singing voice that can express natural vibrato. The study reveals a simplest form of synthesis since, a lyric syllable in a Mandarin song has only a music note assigned to it. The pitch-frequency is the major acoustic factor that the researchers have dealt with. Rather than synthesizing vibrating pitch by applying some rules researches focus on constructing ANN based models. The models initially generate vibrato parameters of corresponding vibrating pitch contour. Then, in terms of these parameters, a vibrato-expressing pitch contour is generated indirectly. Each ANN is a Multi-Layer Perceptron (MLP) and requires high quantity of recorded syllables for training. IPF (instantaneous pitch frequency) curve is measured for syllable signal files and further analyzed to extract intonation, vibrato extent, and vibrato rate parameters. The approach a moving average based filter to re-generate intonation curve.

[19] Is an interesting insight to more naturalistic singing interpretations. Research also mentions on Thayer's two dimensional model of emotion while introducing a Hidden Semi-Markov Models (HSMM) based emotional singing voice synthesis system. The procedure to synthesize singing voice is includes determining the sequence of pitch and duration for a given musical score, sampling duration, spectral parameters and vibrato parameters from appropriate HSMMs and synthesizing singing voice by Mel-Log Spectrum Approximation (MLSA) filter using the sampled features. In the synthesis procedure, duration and vibrato parameters are controlled to express emotions.

[20] Compares on Voice conversion with Maximum-likelihood Gaussian mixture model, voice conversion with weighted frequency warping, Model adaptation in HMM-based Text-To-Speech (TTS) framework and Spectral transformation in speech-to-singing. This paper focuses on the spectral transformation from speech to singing by extending two types of state-of-the-art techniques for singing synthesis and examining their performance with other alternatives. Experiments indicate that the extended transformation with model adaptation on large data offers the best quality and similarity, where music context-specific transformation contributes to the outstanding performance.

vi) Articulatory Synthesis

2-D Physical model based on the digital waveguide mesh is proposed by [21] for the synthesis and analysis of speech and singing. Real data on the shape of the human vocal tract while in use is hard to obtain. Evolution has been shown to be an effective alternative design method for the shape of such models. Realistic sounds are produced through synthesis, even though the synthesis model used is fairly simple. The facts revealed highlight the need of an evolved model to match the actual vocal tract. Research suggested that evolution is an effective design technique not only in this and other applications where we wish to produce an accurate simulation of reality, but also for the synthesis of new sounds not available acoustically. It is also proved that multi-dimensional waveguide mesh could easily simulate a 10-dimensional object but designing the actual shape of such structured object would be extremely difficult.

vii) Sine-wave Synthesis

Further elaborating the singing synthesizing process, LYRICOS [22] is another important study. It is a system which uses data driven methods to model the phonetic information in the voice, resulting in an output that assumes the voice identity characteristics of recorded human vocalist and employs a high quality sinusoidal synthesis method. Voice data corpus used in the system is collected exhaustively by audio singings of 500 non sense words by trained vocalist. This approach provides support in synthesizing more natural singing than synthesized wave forms using algorithms. According to the target specified by the Musical Instrument Digital Interface (MIDI) input, best inventory units are selected and the concatenation is done within the sinusoidal model frame work. These frameworks defined in LYRICOS exemplarily reveal methods to create synthesized singing output with high quality. LYRICOS also use audio chunk concatenation in its sinusoidal approach.

B. Experimental Interfaces to Communicate Musical Information to Visually Impaired

The In the context of visually impaired music learning, one of the best and well known research is Weasel: A System for the Non-visual Presentation of Music Notation [23]. It is one of the most attention grabbing researches. It allows users to navigate through manuscripts with the help of a tactile

hardware interface and an auditory interface. Weasel overlay consists with two main sections referred as bar area and control section. Once when the user presses on the bar area corresponding musical content is played. Circular symbols below the guidelines represent the presence of dynamic information which consists in a manuscript made for sighted people and will be conveyed to the user through synthetic speech. The control section allows the user to select between different modes available. The modes allow users to navigate within the manuscript much more easily.

The research suggests that it is always better to have several output formats rather than restricting to one particular output format. Weasel seems addressing the point in the exact way. Basically Weasel research highlights that with Braille music; all the symbolic and written instructions are translated resulting, again, in large amounts of information presented in a serial fashion. Therefore the research has also considered on some successful solutions in order to overcome the limitations within Braille music representation and talking scores. The survey carried out parallel to the research highlights that existing methodologies used results with large amounts of information presented in a serial fashion which again limit the user to view more information at one glance. It provides a solution for finding out a way to filter unnecessary information. According to the researchers learner's reading speed is directly limited by the Weasel I & Weasel II [23] speed of the spoken description in talking scores even though prerequisite knowledge on Braille music is not needed. And addressing the problem in western music context seems less confused comparing with eastern music context since western music is more structural than eastern. The number of bars-to-a-line, lines-to a-page and graphical representation of the notations makes it much easier in the converting process. The major advantage of such a system is that user can get rid of serial information retrieval and can experience roughly an overall idea (through experiencing content within a music page using an interactive interface) on the music score more easily, precisely and quickly. The findings also suggest that delivering more controllable output is also important. As the designers has concerned more on represent each and every information available in a manuscript notation, no information is lost as part of translation.

As standard intermediary files used to represent music notations, Music XML format and Braille Music XML format are considered as most usable. [24] Discusses points about BMR (Braille Music reader), BME (Braille music editor) and Resonare (Braille page formatter) to facilitate Braille music.

One of the main problems they have faced in addressing the problem is different scores use different rules in music elements in Chinese music. But in most of other music traditions, there is an acceptance of a common interpretation of music elements like pitch, rhythm which does not raise this problem in our research study. And the study is focusing on converting text to another text. The study discusses that the interpretation has two parts rhythm and pitch (two basic characters of a music notation).

One of the key points is that the paper describes is they have used the features of the notation like notes sequence (NS), numbers of notes (NN), pitch interval position and direction (PIDP) and their combinations: NS+NN, NN+PIDP, NS+PIDP, NS+NN+PIDP have used to develop the interpretation model. This is mainly used because the described notation structure does not give the measure of rhythm of the script.

Converting the visual score in to vibro-tactile is another common and effective methodology to give the sense of the score to visually impaired people. In this context, [25] is one of the most attention grabbing research that was being conducted. It is a well-known factor that if the visual score is being converted in to vibration then to sense the output it is obviously requires specific hardware. The research [25] suggests a mechanism to read the music score with vibrations and as an extra feature Vibro-tactile Score Editor (Vib-Score Editor) that fully supports the vibro-tactile score with intuitive graphical user interface is also being developed. Vibro-tactile is being considered as one of the most efficient HCI mechanism and the researches of [25] tested the potential of common VibeTonz system and vibrotactile actuators such as piezoelectric actuators.

Apart from that the software tools like Hapticon Editor, Haptic Icon Prototyper, VibeTonz studio, and posVibEditor is being considered. Apart from that the research uses XML to store data. As both [23] and [25] suggest, if we utilize the vibro-tactile output format obviously the learning curve decreases since visually impaired people already practiced to read things by tactile (Braille reading). On the other hand the output format becomes comprehensive compared with audio format but time consuming. I.e. giving two different qualities of the sound simultaneously with vibro-tactile is bit confusing. But in audio output format it is not that confused.

As [26] concludes, Vib-ScoreEditor has another metaphorical feature, vibrotactile clef (analogous to the musical clef), to make the process of composing vibrotactile patterns decoupled from the process of considering the signal-level characteristics of the vibration. This suggests that there can be some basic level music symbols that can be define a corresponding vibration signal and there can be some other which cannot be. And the mechanism to be followed to generate vibro-tactile output is confusing, as both [25] and [26] suggest.

[27] Study presents a detailed description of an evaluation carried out of a vibro-tactile pattern design using vibro-tactile score. Apart from that the research [27] suggests two basic mechanisms named "Wave form editing" and "Score editing". Though the output format is a vibration, they first generate a wave form of that vibration signal.

Converting visual script to thermo-tactile output is another innovative method of presenting music notation to visually impaired people. In this discussed context there were significant researches carried out. The study [28] is one of such researches which attempts to develop a thermo-score display dynamically changes the temperature of the instrument according to the frequency of the notes. The MIDI signals were converted to temperature using MIDI-to-temperature converters. To generate and control the temperature researchers have used Peltier device. And the device is placed on the

instrument where the performer get the music signals through the instrument itself and using those bio-signals the performer generates the music with the instrument. And the system discussed by [28] basically controls the time which performer plays a note. The system is mainly considered about keyboard playing and making one key hotter than the temperature the performer leads to pull finger by reflex action which makes the note plays short time. The research is important to us since it develops a non-visual output methodology. Basically this system is generated to use in instrumental performance in music which always requires the music instrument to sense the music notes. The study highlights an important point that is it claims a performer always conducts music information processing which implies the feedback the performer gets both from the audience and the instrument affects to the performance.

The instrument discussed in the study make the frequently used keys hotter than others. In order to do that it first creates a chroma-profile (chart representation of frequency of pitch notations computed from MIDI data). The paper discusses further more details that can be given to the performer using the device. It suggests with the same device giving the audiences excitement. It depends on the performer. Since there can be performers whose performance may be less when user excitement goes up, if so even the average performance even may not achieved as expected through the device.

Conceptually a music trainer can be defined as one who gives a genuine feedback on a musical performance by analyzing it. According to Study on Software-Based Extraction of Objective Parameters from Music Performances by Alexander Lerch [29] Music Performance Analysis (MPA) aims at studying the performance of a musical score rather than the score itself. It deals with the observation, extraction, description, interpretation and modeling of music performance parameters as well as the analysis of attributes and characteristics of the generation and perception of music performance. According to him four classes of acoustical parameters that can be used for the description or characterization of music performances have been identified. They are as [29]. Tempo / timing (global or local tempo and its variation, rubato, or expressive timing, subtle variation of note lengths in phrases, articulation of tones, etc.), Velocity, loudness or intensity (musical dynamics, crescendo and diminuendo, accents, tremolo, etc.), Pitch (temperament, tuning frequency, expressive intonation, vibrato, glissando, etc.) and Timbre (sound quality and its variation resulting from instrumentation and instrument specific properties such as bow positioning). Various researches has gone through experimenting various methods in extracting above parameters from a music performance, analyzing them in order to come up with conclusions on the performance, filtering out separate performances from a polyphonic music, discriminating between vocals and non-vocals etc.

One of the most intriguing findings on auditory display and sonification are brought forward through recent [30]. The goal of Auditory Display is to enable a better understanding, or an appreciation, of changes and structures in the data that underlie the display. Auditory Display encompasses all aspects of a human-machine interaction system, including the setup, speakers or headphones, modes of interaction with the display system, and any technical solution for the gathering, processing, and computing necessary to obtain sound in response to the data. In contrast, Sonification is a core component of an auditory display: the technique of rendering sound in response to data and interactions [30]. It proves that there is immense amount of information that humans are able to extract from sounds. It also acknowledges that none of the attempts to use sounds in synthetic auditory displays has yet come close to conveying that amount of information.

Tangible User Interfaces (TUIs) which are still at research stages also will play a major role in assistive technologies for visually impaired and also for people with multiple disabilities, in the years to come.

Specially in communicating composite signals, 3D sound models contain more successful potential. Researches prove that 3D sound models can be easily used as a media to convey multiple sound information, if there is only less number of varieties. Knowles Electronics Manikin for Acoustic Research (KEMAR) is mostly used for evaluation of such models which depends on head-related transfer function (HRTF).

III. Modelling Shape of a Sound

All the knowledge paths unlocked by the key literature allow us to come up with a new approach to model synthesized singing. Hybrid techniques which basically depend on concatenative synthesis are used in this method. A concept which models a shape of a musical sound is brought forward and operators are defined to distort, merge, and transform the model depending on the acoustic parameters such as timbre, pitch, rhythm, intensity and timing. One or more sound shapes are concatenated to make a phoneme with specific frequency.

The problem of most of the existing singing synthesizers is that most of them fail to provide a naturalistic output and naturalistic effect when tempo has been changed. Most of the signals are synthesized from the scratch or even recorded audios used the functions executed on them results in a violated output far away from expected natural output. The model in this study is to get the use of recorded audio as far as we can and re-defining functions to produce more naturalistic output which is closer to the expected. Concatenative synthesis is regarded as the most successive singing synthesizing method. To get bonded to a specific scope, the model is presented to Sinhala language and for eastern music notation scripts. Sinhala is a strong language which has skillfully filtered out the units of sound which human voice can generate. Therefore, there is a huge chance that using phonemes in Sinhala language will make the synthesizing task more flexible. Eastern music has inherited every type of craftiness which can be embedded within singing. Therefore effect of Kan swara, Blend and Meand also has brought into focus.

The definition of this conceptual models involved thorough analysis to detect how the audio wave forms behave when the frequency and tempo changed relevant to a specific phoneme. And on how position in the syllable and neighboring phones are affected the wave form. Initially, these scenarios are tested relevant to one phoneme (like the aalaap).

Overall synthesizing model is depicted in Figure 1. Notation sequence and lyrics sequence is used to come up with the general sound shape model for a specific phoneme. Depending on the derivative parameters, the base sound model is re-shaped accordingly. Depending on the thaal sequence which defines the required tempo, shape is shrinked or stretched to fit to a single time unit (maatra). Base shape model is used to retrieve the basic sound clip from a sound bank (audio database) and the model is then converted from shape domain to frequency domain and applied on top of the audio chunk retrieved. Finalized auditory unit output will be then forwarded to sequencer in order to align on the time line accordingly.

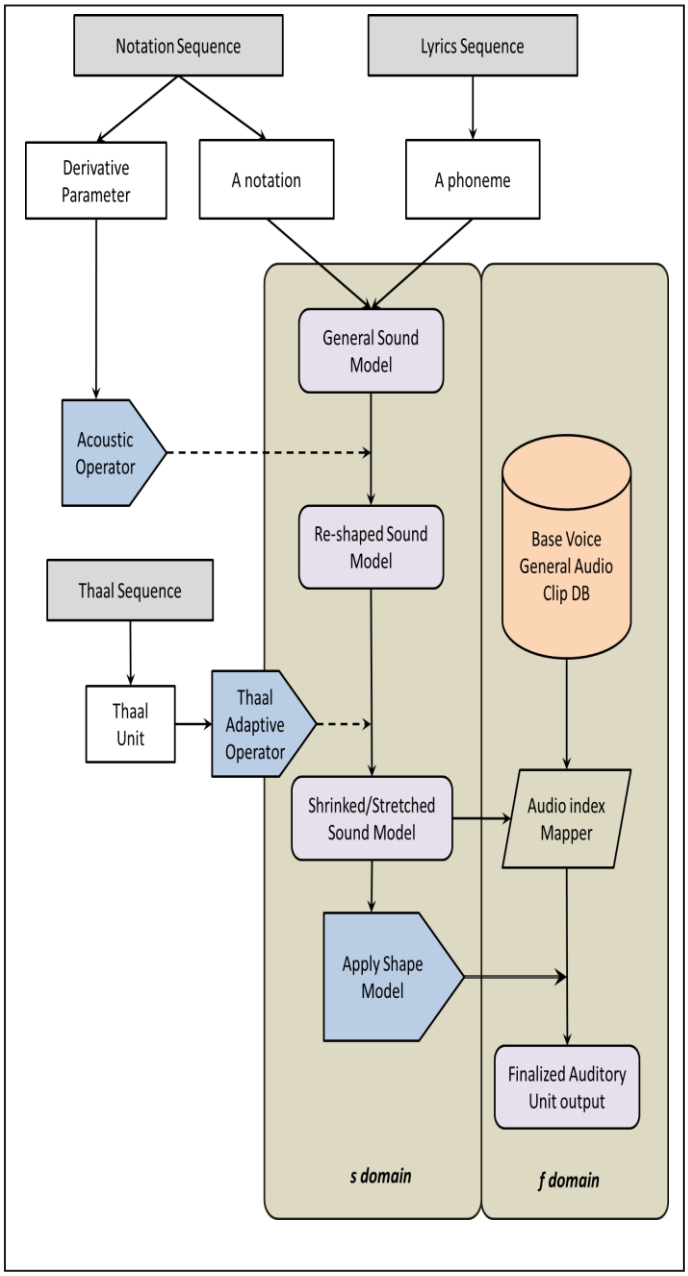

Figure 1. Overall process flow

IV. Evaluation of the Model

Evaluation of the model needs to be done in both shape and frequency domain. Expected shape created from a recorded audio chunked needed to be converted to its shape and frequency. The models need to be compared with the actual outcome. This need to be followed to evaluate the outcome of a series of audio units and also for single units as mentioned above since the sound unit shape deviates from the output of a sound series.

V. Conclusion

When considering the outcome of the researches, concatenative synthesis is considered more reliable and effective method which can generate more naturalistic output. With this type of a model it is possible to define a universal language to represent each type of sound shapes. Through the synthesized singing output which is capable to adapt for any tempo visually impaired users will have independent access to majority of eastern music notation scripts together with lyrics. It will unlock paths to a singing synthesizer specifically adapted to Sinhala Language. Users also will be able to use the solution to access both lyrics and notes even in a new composition. Users will be delivered more control in navigating through the notation script when relevant information layers were kept in enabled or disabled mode to get rid of information overloading.

Synthesizing capabilities especially for emotional singing are aimed to be developed in future studies and also the capability to base upon many male and female base voice types. Main future target is to develop solution up to an assistive tool which can be used in music composing and learning with interactive and user friendly interfaces.

The outcome of the research will fill the existing accessibility gap which visually impaired music lovers face and is expected to enhance the opportunities they have to study and experiment on music independently.

Acknowledgment

Late Prof. D. P. M. Weerakkody, Faculty of Arts, University of Peradeniya and Ms. Shantha Shyamalee, Eastern music teacher, De Soysa Maha Vidyalaya, Moratuwa are reminded with respect and gratefulness for their guidance and support given. Without the great support of students of School for the Blind, the Ceylon School for the Deaf and Blind, Ratmalana it might be very hard to carry out the surveys and usability testing in order to carry out successful research evaluation. Hence, they are also gratefully reminded. Our heartfelt gratitude also goes to Mr. Amila Indrajith, Music teacher, School for the Blind, The Ceylon School for the Deaf and Blind, Ratmalana, Ms. Padmakumari Karunatilake, Mr. Sarath Kumara IT teachers, School for the Blind, The Ceylon School for the Deaf and Blind, Ratmalana, Mr. Asoka Weerawardena, IT Instructor for the visually impaired, University of Colombo and Ms. Aruni Samarasinghe, Eastern

music teacher, Kanuwana Vidyalaya, Ja-Ela, Mr. B. R. Dassanayake, Dean, Faculty of Music, University of the Visual and Performing Arts, Ms. Sandya Kodduruarachi, the principal - School for the Blind, the Ceylon School for the Deaf and Blind, Ratmalana, Ms. Indu Bandara, Director, Census Division, Department of Census and Statistics and Ms. Mekhala Gamage, famous vocalist.

REFERENCES

[1] D. B. Kiriella, K. C. Ranasinghe, S. C. Kumari, K. L. Jayaratne, "Music Training Interface for Visually Impaired through a Novel Approach to Optical Music Recognition." Journal on Computing (JoC) 3.4 (2014)

[2] Kenmochi, Hideki, and Hayato Ohshita. "VOCALOID-commercial singing synthesizer based on sample concatenation." INTERSPEECH. Vol. 2007.

[3] Huang, Xuedong, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, M. Plumpe. "Recent improvements on Microsoft's trainable text-to-speech system-Whistler." Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. Vol. 2. IEEE, 1997.

[4] C. Chafe, Glitch Free FM Vocal Synthesis, "Center for Computer Research in Music and Acoustics," Stanford University.

[5] Kyritsi, Varvara, A. Georgaki, and G. Kouroupetroglou. "A score-to-singing voice synthesis system for the Greek language." in Proceedings of the Inter. Computer Music Conference (ICMC07), Copenhagen, Denmark, 2007.

[6] H. Kaprykowsky and X. Rodet, "Musical Alignment Using Globally Optimal Short-Time Dynamic Time Warping Short-Time Dynamic Time Warping.".

[7] H. Kaprykowsky, and X. Rodet, "Globally optimal short-time dynamic time warping, application to score to audio alignment." Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference.

[8] E. Molina, I. Barbancho, A. M. Barbancho, L. J. Tard, "Parametric Model Of Spectral Envelope To Synthesize Realistic Intensity Variations In Singing Voice," ATIC Research Group.

[9] P. Y. Chan, M. Dong, S. W. Lee, L. Cenychan, mhdong, swylee, "Solo To A Capella Conversion - Synthesizing Vocal Harmony From Lead Vocals"

[10] J. Janer, J. Bonada, M. Blaauw. "Performance-Driven Control For Sample-Based Singing Voice Synthesis," 9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada, September 18-20, 2006.

[11] K. Ota, T. Ehara, "Four-tone modeling for natural singing synthesis in Chinese and comparing synthesized singings with speaking voices," Proceedings of 20th International Congress on Acoustics, ICA 2010. 23–27 August 2010, Sydney, Australia.

[12] T. Saitou, M. Unoki, M. Akagi. "Development of the F0 Control Model for Singing-Voices Synthesis," School of Information Science, Japan Advanced Institute of Science and Technology.

[13] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-To-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 21-24, 2007, New Paltz, NY

[14] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing Voice Conversion Method Based on Many-to-Many Eigenvoice Conversion and Training Data Generation Using a Singing-to-Singing Synthesis System," Graduate School of Information Science, Nara Institute of Science and Technology (NAIST).

[15] Y. E. Kim, "A framework for parametric singing voice analysis/synthesis." In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on., pp. 123-126. IEEE, 2003.

[16] N. Ramanathan, "Applications of Hidden Markov Models," 2006.

[17] R. Li, Y. Ding, W. Li, and M. Bi, "Automatic Interpretation of Chinese Traditional Musical Notation Using Conditional Random Field," no. 60933004, pp. 19–22.

[18] H. Gu, Z. Lin, "Mandarin Singing Voice Synthesis Usingann Vibrato Parameter Models." Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology.

[19] Y. Park, S. Yun and C. D. Yoo. "Parametric Emotional Singing Voice Synthesis," 2106, 305-701, Republic of Korea.

[20] S. W. Lee, Z. Wu, M. Dong, X. Tian, and H. Li, "A Comparative Study of Spectral Transformation Techniques for Singing Voice Synthesis," Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore.

[21] Cooper, C., Murphy, D., Howard, D. and Tyrrell, A., "Singing synthesis with an evolved physical model," IEEE Transactions on Audio, Speech and Language Processing, Volume 14 (4), 1454 – 1461, 2006.

[22] M. W. Macon, L. J. Link, J. Oliverio, M. A. Clements, E. B. George, "Concatenation Based MIDI to Singing Voice Synthesis," in Audio Engineering Society Convention 103. Audio Engineering Society, 1997..

[23] B. P. Challis, and A. D. N. Edwards, "Weasel: A System for the Non-visual Presentation of Music Notation," Computers Helping People with Special Needs: Proceedings of ICCHP 2000, pp. 113-120, Karlsruhe, Germany, Osterreichische Computer Gesellschaft.

[24] G. Bertoni, G. Nicotra, and A. Quatraro, "Access to music by blind people."

[25] J. Lee, J. Ryu, and S. Choi, "Vibrotactile score: A score metaphor for designing vibrotactile patterns." EuroHaptics conference, 2009 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics 2009. Third Joint. IEEE, 2009.

[26] J. Lee, J. Ryu, and S. Choi. "Graphical authoring tools for vibrotactile patterns." EuroHaptics conference, 2009 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics 2009. Third Joint. IEEE, 2009.

[27] J. Lee and S. Choi. "Evaluation of vibrotactile pattern design using vibrotactile score." Haptics Symposium (HAPTICS), 2012 IEEE. IEEE, 2012.

[28] Miyashita, Homei, and K. Nishimoto. "Developing a Non-visual Output Device for Musical Performers." Proc. Sound and Music Computing 4 (2004): pp. 251-255.

[29] A. Lerch, "Software-Based Extraction of Objective Parameters from Music Performances," 2009.

[30] T. Hermann, A. Hunt, J. G. Neuhoff. "The sonification handbook." Berlin, GE: Logos Verlag, 2011.

Mr. Kavindu Ranasinghe obtained his B. Sc. in Information and Communication Technology from University of Colombo School of Computing (UCSC), Sri Lanka. His research interests include Image Processing and Computer Vision, Audio Signal Processing, Character Recognition, Human Computer Interaction, Multimedia Computing and Music Information Retrievals.

Dr. Lakshman Jayaratne - (Ph.D (UWS), B.Sc.(SL), MACS, MCS(SL), and MIEEE) obtained his B.Sc (Hons) in Computer Science from the University of Colombo (UCSC), Sri Lanka in 1992. He obtained his PhD degree in Information Technology in 2006 from the University of Western Sydney, Sydney, Australia. He is working as a Senior Lecturer at the UCSC, University of Colombo. He was the President of the IEEE Chapter of Sri Lankan in 2012. He has wide experience in actively engaging in IT consultancies for public and private sector organizations in Sri Lanka. He was worked as a Research Advisor to Ministry of Defense, Sri Lanka. He Awarded in Recognition of Excellence in Research in the year 2013 at Postgraduate Convocation of University of Colombo, Sri Lanka. His research interest includes Multimedia Information Management, Multimedia Databases, Intelligent Human-Web Interaction, Web Information Management and Retrieval, and Web Search Optimization. Also his research interest includes Audio Music Monitoring for Radio Broadcasting and Computational Approach to Train on Music Notations for Visually Impaired in Sri Lanka.