ABSTRACT
This paper presents an interface involving five degree-of-freedom (DOF) mechanical control of a two-dimensional, mid-sagittal human tongue-like structure for articulatory speech synthesis, enabling user to interactively generate varieties of vocal sounds. As a demonstration of the project, the user will learn to produce a range of sounds, by varying the shape and position of the upper surface of the tongue-like structure in 2D space through a set of three sliders mounted on movable platform. A spring based tongue-like structure and modeling-clay based fixed upper palate form a simplistic physical representation of the anterior part of upper airway. In the sound synthesis engine, the corresponding acoustic model is derived and coupled to a simple two-mass glottal model that serves as the excitation source and a microphone is placed in front of the lips which act as the listener. The magnitude and frequency of the glottal excitation is controlled physically by two additional sliders. This entire arrangement allows the user to play around with five sliders to vary articulatory structures as well as the source acoustic parameters, exploring the variation of sounds.

KEYWORDS
five degree-of-freedom control, physical interface, sound synthesis engine, sliders, dual-handed simultaneous control

1 INTRODUCTION
Articulatory speech synthesis is currently one of the most challenging research domains, that encompasses the production of speech sounds using a vocal tract model and simulating the movements of the speech articulators like tongue, lips, velum etc. Such synthesizers are utmost important for understanding the mechanism of human speech production and study of phonetics. Besides, it provides great insights to the acoustic aspects of human speech.

Many musical and audio interfaces such as Digital Duets [7], MIDI etc have already been designed and implemented which mimic the actual acoustic phenomenon of real musical instruments. However, despite having considerable significances in research and learning purposes, there is a dearth of intuitive user interfaces to effectively control and synthesize intelligible articulatory sounds based on simultaneous variation of speech articulators.

This work implements a 2D mechanically controlled tongue-like (closely relating to tongue properties) structure with a novel five DOF control scheme targeting vocal sound synthesis. The goal is to develop a convenient, easy-to-learn and cost-effective physical interface with improved mechanical control leveraging multi DOF capability of human arm. We use a simple spiral spring to model the upper surface of tongue and cylindrical clay wrapped with black tape to represent the fixed upper palate and lingual base (upper and lower boundaries). This arrangement enables the user to utilize one of his/her hand to vary a set of sliders and a mobile platform with the sliders mounted on it, to modify the shape of tongue surface and its position in 2D plane. This alters the anterior part of the upper airway, thereby modulating sound propagation through it. Additionally, we provide a two-way (slider and mouse based) control of the gain and frequency of source sound (glottal excitation), to investigate user preference in controlling acoustic parameters, through the other hand.

The proposed interface has been compared on the basis of convenience, expressibility and learnability aspects of control with two popular software applications, which will be detailed in the next section. The users are allotted some specified tasks that attempt to answer our research question; the results are evaluated both quantitatively as well as qualitatively and finally interpreted extensively, to investigate the usability of the proposed interface.

2 RELATED WORKS
2.1 Pink Trombone
Pink Trombone [4] is an online voice synthesizer application that presents an interactive mid sagittal view of human vocal tract, which can be manipulated by users through mouse control to simulate various vocal sounds. The users can slide the variable circular purple tongue, lips, hard palate and velum, and consequently hear the vocal sounds in real time, ranging from shrill screams to low
rumbles. It is an integration of vocal tract, nasal tract and the glottis - from which sound is generated.

The tongue position can be changed through mouse, manipulating a circular point moving freely in a triangular control space. The cursor can be dragged over the tongue surface and held fixed at a particular point, clicking the left button of the mouse to change the tongue shape to that particular shape, but as soon as the left click is released, the tongue shape changes consequently. So there is no intuitive arrangement to lock the tongue shape to that particular shape, which is a major shortcoming of this interface. The glottal excitation can be varied through separate set of controls known as Voice-Box Control which performs the pitch (frequency) and gain variation. There are two discrete coarse levels of gain. Moreover, there is another option for continuous fine tuning of gain and frequency through same slider. In other words, increase (or decrease) of frequency and gain occur simultaneously, and it is not possible for the user to continuously change gain and frequency in opposite directions. This is another major control issue.

In addition to that, it has a few control glitches. The application gets easily corrupted when user drags the cursor, with simultaneous right-clicking the mouse, as shown in Fig 1.

2.2 VTDemo

VTDemo [1] is a Windows PC interface that investigates the variation of sound quality corresponding to changes in vocal tract shape. The interactive application gives the user an opportunity to change the sliders mapped to several vocal tract and glottal excitation parameters and thereby synthesize vocal sounds in real-time. The sound synthesis engine relies on an area function based one dimensional wave equation, similar to Pink Trombone, to generate the sounds.

The control panel has a total of 10 sliders, aimed at changing the Jaw Height (JW), Tongue Position (TP), Tongue Shape (TS), Tongue Apex (TS), Lip Area (LA), Lip Protrusion (LP), Larynx Height (LH), Glottal Area (GA), Fundamental Frequency (FX) and Velo-pharyngeal port opening (NS) individually, as shown in Fig 2. The sliders can be varied only one at a time, implying that the user has to change the parameters one after another to achieve a target vocal tract shape and target sound. Correspondingly, the animated digital vocal tract undergoes these changes one at a time independent of each other. However, in real world, the human vocal tract articulators work simultaneously in an extremely interdependent manner, due to the intermingling of muscles, bringing about the articulatory movements. Hence, this kind of control becomes highly unrealistic when compared to articulatory speech production.

For this work, we are mostly concerned about the control of tongue surface, involving sliders TP, TS and TA and the glottal parameters involving GA, FX. The tongue is a highly deformable, muscular hydrostat organ with infinite DOFs, equipped with eleven muscles (extrinsic and intrinsic) controlling the shape and position of tongue. So evidently, only three sliders with range [-3,3] are not sufficient to map the slider inputs for changing the shape and position of tongue. These changes occur in some predefined trajectories which are less intuitive and difficult to relate to the slider changes. There is a lack of user flexibility, since a user has to select a particular shape among a number of pre-defined tongue shapes, corresponding to changes in slider values.

3 NOVELTY AND CONTRIBUTION

The tongue plays the most significant role in varying the vocal tract structure, responsible for generating wide variety of sounds. However, most of the available interfaces aiming articulatory speech synthesis are focused towards the sound synthesis mechanism through better, realistic modeling of articulatory elements. Hence, more research needs to be directed towards the actual user-interface facilitating the control and manipulation of the tract contour. So, this work, besides performing one dimensional sound synthesis based on 2D vocal tract shape variation, aims at developing a better physical user-interface through multi-degree of freedom control arrangement. Human vocal tract and tongue, in reality, being extremely dynamic and flexible, perform multiple position and
shape changes simultaneously. This smooth and efficient coordination between multiple articulators with infinite DOF leads to the actual production of vocal sounds.

The conventional user interfaces investigating the issue like Pink Trombone, VTDemo employ mouse based one-point control which allow the user to control one slider or one point on the vocal tract at a time. It essentially enables user to explore the effect of only one acoustic parameter or shape/deformation of one part of the tongue for production of vocal sound. The other parts of the same articulator or different articulators are assumed to be fixed, which might be one of the reasons behind lack of realistic quality in vocal sound synthesis. Besides, since it is only a single point control, the vocal transitions (like, from /a/ to /u/) are very slow and involve intermediate unwanted sounds. This is because one (or two) DOF control does not allow the smooth articulatory trajectories necessary for the production of speech.

In this work, we attempt to address these issues of simultaneous, multiple DOF control by allowing the user to manipulate the shape as well as position of the tongue surface through both of his/her hands. We develop a simplistic partial representation of the functionality intrinsic and extrinsic tongue muscles through three sliders mounted over a movable platform. The sliders partially serve the purpose of intrinsic tongue muscle, allowing the user to vary tongue deformation. The platform moves freely in parallel and transverse direction on a bounded plane to vary the tongue position, thereby somewhat serving the purpose of extrinsic tongue muscles. So instead of utilizing a direct control used by the available sound synthesis applications, here, we tend to involve a sort of force-based position control to allow simultaneous five DOF movement i.e two DOF for tongue shape variation and three for the tongue deformation at tip, top and base, which supposedly increases the ‘controllability’ of the interface. Besides, most of the interfaces allow mere independent controls of various parts, which means control of one part of the articulator do not reflect any changes in the other parts or do not provide any feedback to the user for the variations in other interrelated parts. However, in reality, our muscles and articulators are intimately interleaved, because of which, movement in one part of an articulator renders changes in other parts as well. Therefore, we choose a flexible spring structure to replicate the tongue surface which allows user to change the entire tongue surface deformations with these three control points. As a result, if the user decides to control only one part of the tongue through one slider control, simultaneous angular movement is noticed in the connecting links for the other sliders, which enhances the ‘expressibility’ of the design. In other words, the user receives a visual feedback regarding the interdependency of different parts of the same articulator, through only following the angular position of the other slider-connections. This provides the notion of an added affordance to the user, which is vital in developing musical interfaces, but is lacking in conventional articulatory speech synthesis applications.

Moreover, the physical interface can give the users an idea regarding the purpose of the articulatory muscles, kinematics behind articulatory movements and generation of speech sounds through such movements. By getting trained through this physical model, users can learn to control an artificial tongue through his hands, to get a second voice. Thus, we are interested in exploring the learnability of our control as well.

4 MECHANICAL INTERFACE

In order to build the experimental apparatus, we use a cork board as the base, card boards for control space and movable platform, Arduino, document camera, slider sensors, mouse, a laptop and a speaker. To create a sagittal view of the tongue surface on a 2D plane, a spiral spring element is used and three different points of it (tip, top and root) are connected to the three sliders through thin aluminium rods, as shown in Fig 3. The user would be able to vary the shape and the position of the spiral element within an articulatory space bounded by the upper palate and the lingual base. The hardware interface is divided into two controlling blocks. The primary one is targeted to change the tongue shape and position while the secondary one is aimed to control the source (glottal) frequency and gain. The user would be able to use both the blocks simultaneously through an ambidextrous (dual-handed) control scheme. The overall latency of the interface was found to be 0.08 seconds.

4.1 System Design

4.1.1 Primary Block: Tongue Shape and Position control. Three sliders are mounted over a movable platform and attached to the spring to control and change the tongue shape as shown in Fig. 4. Keeping the platform fixed, users can vary the slider positions with
three fingers to create desired tongue shapes. Similarly, to change
the tongue position separately, the user can move the platform on
a glossy cardboard base with specified boundaries, as shown in
Fig 5. To restrict the unrealistic angular rotation of the tongue-like
structure, the movable controller is enabled with only two DOF,
by the arrangement as demonstrated in Fig 4, which allows its
displacement along x and y direction only. This design provides the
five DOF, which users can control simultaneously with one hand
ideally, as evident from Fig 3.

4.1.2 Secondary Block: Source (Glottal) frequency and gain con-
trol. As the primary block involves more complexity, to simulate
this functionality we came up with two designing ideas which could
reduce the user’s effort while controlling both the blocks simulta-
neously. In the first design, we use two slider sensors connected to
a microcontroller and the position of both the sliders are tracked
to vary the source frequency and gain, as demonstrated in Fig 6. 
In the alternative second design, we use a mouse connected to the
laptop which can be moved within a predefined window so that
the cursor position can be tracked and mapped accordingly, to
a gain-frequency co-ordinate system, as demonstrated in Fig 7. With
the left click button pressed, moving the mouse towards right (or
left) would correspondingly increase (or decrease) the frequency
and moving down (or up) would correspondingly increase (or de-
crease) the gain. The user can easily follow any desired trajectory
in any direction, in order to sequentially vary the glottal frequency
and gain.

The whole set-up is placed under a document camera as shown
in Fig 3, and the camera is connected to the PC for real-time pro-
cessing.

4.2 Real-time Area function Computation

The real-time video captured by the document camera is read
through Image-Mate Software, in MATLAB environment. The cor-
responding image frames are extracted at a frame rate of 30 fps.
The rigid upper palate and tongue-like spring structure is detected
and extracted utilizing three distinct variations in vertical image
intensity profiles. At regular spatial intervals i.e 30 control points (i)
along the structures, we compute the vertical distance (di) between
the palate and tongue surface or the lower boundary, as shown
with Red arrows in Fig 7, and using these values as the diameters,
we derive the corresponding 2D area functional values (Ai) [2] that
will be utilized for the next step.

4.3 Sound Synthesis Engine

The goal of this design is to approximate model the vocal tract
sound propagation by using a waveguide model in 1D acoustical
tube. For sound propagation in vocal tract, a well known physi-
cal model is Kelly-Lochbaum (KL) which employs a 1D acoustical
tube structure characterized by an area function. The underlying
idea of KL model is that a 1D plane wave that surfaces from far
end of the vocal tract (glottis) travels through a line of concen-
tric cylinder segments with varying cross-sectional areas defined
by area function to the open mouth end. In this work, we imple-
ment a method described in [5] which eliminates the drawbacks of
KL model. The vocal tract is modeled as an acoustic tube with its
shape changing accordingly with the area functions received from
image processing module. Glottal excitation pulse was generated
according to the Rosenberg’s model [3]. This vocal fold model is
coupled to discretized acoustic equations in the vocal tract. The
acoustic wave propagation is simulated by numerically integrating
the linearized 1D Navier-Stokes pressure-velocity PDE in time and
space on a non-uniform grid. The synthesis mechanism involves
excitations acting as source placed in the tube and sound prop-
agation being simulated by approximating the pressure-velocity
wave equations[5]. All these models are implemented in Java Audio
5 USER EVALUATION

5.1 Participants
A total of 13 participants (8 females and 5 males, in an age range of 20-34) were selected from UBC. Prior to the user study, we conducted a pilot test of our prototype with 3 participants, which helped in designing our user-study and address the technical shortcomings of physical interface. The data collected during the pilot testing was discarded. We selected a wide range of participants based on their familiarity on the principle of articulatory sound synthesis, as shown in Fig 9. Before beginning the experiments, participants gave written consent for the investigators to use the survey responses and audio recording of the interview.

5.2 Experimental Procedure
As the experimental focus is primarily based on how the users control a physical interface of the human vocal tract (mostly, the tongue-like structure), we eliminated the vocal sound generation as a measured variable. And to avoid any bias for that, we take great care during the experimental design. Each participant goes through a forty minutes session that consist of five phases, starting with a short introductory session for every participant to have a basic knowledge regarding articulatory sound synthesis. As we compare the controlling aspect of our physical interface (Sound Stream) with the VTDemo (desktop application) and Pink Trombone (web application), next, we conduct an initial training session for the users across all the three interfaces before assigning the user study tasks.

5.3 Task Design
We design six set of tasks to validate and compare the controlling elements of the Sound Stream interface against the other two interfaces. The first three tasks are designed to measure the user’s preferences while controlling multiple degree of freedoms to change the tongue shape and the position simultaneously. Though the source (glottal) frequency and gain are not related to the tongue shape and position change, it plays an important role while producing vocal sound with various pitch and loudness. So, to give more control to the users over the sound and to make a fair comparison with other interfaces we ask the user to use both the slider and a mouse to understand their flexibility and convenience in using a both-handed control. The number of degree of freedoms involved, are increased with the tasks to increase the difficulty levels of task. First, the participants are asked to change the tongue shape with three sliders; followed by shape and position change through controlling all 3 sliders and the movable platform; and finally tongue structure change (shape and position) along with acoustic parameter variation (gain and frequency). The last three set of tasks are designed to verify how the Sound Stream interface facilitates the users to create different tongue shapes with minimal effort. So, we predefine three tongue structures based on their complexity (simple to high) using all the three interfaces. And participants are asked to create the same tongue shapes within a limited time to measure the error. Since these interfaces have different control mechanisms and the interfaces are widely varying, it is difficult to make similar tongue shapes in three interfaces. Besides, the same shape may be easy to attain in one interface and difficult in the other. We resolve this issue, by specifying only minor shape change from base shape and position for ’simple’ structure (one slider change in VTDemo and Sound Stream, 2 unit movement of control point in Pink Trombone), major shape change (use of TA, TS sliders in VTDemo and all three in Sound Stream, 4 unit movement of control point in Pink Trombone) for ’medium’ and simultaneous shape and position change (use of TS, TA, TP sliders in VTDemo, all 3 sliders and moving platform in Sound Stream, 5 unit movement of circular control point along with cursor-drag along the tongue surface in Pink Trombone) for ’complex’ structure. We reset the structure before starting individual tasks each time.

5.4 Questionnaire and Interview
After completing the tasks, the user has to fill out a Google questionnaire form. The questionnaire is designed to collect the demographic data of the participants and their feedbacks in a scalable format for quantitative analysis. Following this, an interview session is arranged to have more insights into the challenges that the users went through, while interacting with all the three interfaces. Both the quantitative and qualitative analysis were carried out for the user study.
We design three hypotheses to perform the quantitative analysis of the data that has been collected during the user study.

5.5.1 Hypothesis 1: 
H0: There will not be any significant difference in user preferences based on whether a user controls three DOF, five DOF or ambidextrous (AMB) i.e five DOF + Gain and Frequency, simultaneously in Sound Stream.
H1: There will be a significant difference in user preferences based on how many degree of freedom they control simultaneously in Sound Stream.

5.5.2 Hypothesis 2: 
H0: Users will not find any differences in terms of complexity while controlling multiple parameters simultaneously across all the three interfaces (VT Demo, Sound Stream, Pink Trombone).
H1: Users will find significant differences in terms of complexity while controlling multiple parameters simultaneously across all the three interfaces (VT Demo, Sound Stream, Pink Trombone).

5.5.3 Hypothesis 3: 
H0: There would not be any noticeable difference in the error ratios across all the three interfaces while creating a predefined tongue shape under a limited time.
H1: There will be a significant difference in the error ratios among the three interfaces while creating a predefined tongue shape under a limited time.

5.6 Experimental Limitation
The Sound stream interface is a physical interface whereas the Pink Trombone and VT Demo, both are the digital interfaces. And all the three interfaces have a different controlling modality and underlying mechanism of control, though they share a common goal. With our current interface design, there are only three sliders for changing the tongue shape. So, in the future work, if we increase the number of controlling points to get a more accurate tongue shape, then the current experimental result may not pertain. Moreover, it would be better if we could compare the Sound stream interface with an existing physical interface so that the experiment results will guide towards a better design of physical-biomechanical control and sound synthesis engine. However, such hardware interfaces that would allow a multi DOF control for articulatory movements, were not available to us and hence we decided to continue with the digital interfaces for comparison. Also, such a comparison would help us understand the preferences of users choosing software and hardware interfaces for serving the predefined purpose.

However, as we have tested the controlling aspects of our interface with 13 participants only, it is still very hard to generalize the experimental results for all the users. Nevertheless, it helps us to move towards a direction that would increase user convenience in controlling articulatory parameters.

A major assumption in our interface is that, we are neglecting the sound quality as a measure of performance metric, though it is the ultimate aim of interfaces targeting articulatory speech synthesis. Though the user will be able to identify sound variations with changes in tongue shape and positions, the vocal sounds lack clarity and intelligibility. This is because we have used a simple one dimensional wave model to synthesize the sounds, since our main focus is to explore user convenience in a simultaneous control, and not the assessment of the audio generated as a result of the control.

6 RESULTS
6.1 Quantitative Results
A total of 3 repeated measures ANOVA tests are used to analyze the data and validate them against our hypothesis. The first two tests are focused towards the measurement of the controlling aspects of the proposed physical interface while the last test is to measure the error ratio across all the three interfaces.

The first test is targeted to compare whether the users are comfortable enough with ambidextrous controllers having multiple degrees of freedom. There is no significant difference in the average value of participants’ preferences (MVT Demo = 3.84, SDVT Demo = 0.68, MPink Trombone = 3.53, SDPink Trombone = 1.19; MAMB = 3.38, SDAMB = 1.04). And the variance analysis shows the following result: F(2,12) = 1.23 (Fcrīt = 3.40), p = 0.31 (> 0.05), 2 = 1.43.

The second test is supposed to evaluate which interface facilitates better opportunities to control the multiple degrees of freedom with ambidextrous controllers. We run a statistical analysis on the user’s preferences on a scale of 1 to 5 for each interface. The average value of the user’s preference across all the three interfaces are MVT Demo = 2.23, MSound Stream = 3.76, MPink Trombone = 3.15. And the variance analysis shows the following result: F(2,12) = 10.16 (Fcrīt = 3.40), p = 0.0006 (< 0.05), 2 = 15.58.

To test against the third hypothesis, we compute the differences in curvature metrics and positions of tongue tip, top and root points of the target images with the corresponding tongue shape created.
by the users using all the three interfaces for each task level. The computed differences are averaged and considered as the error ratio across all the three interfaces. After the statistical analysis of the error ratios for all the three task levels (simple, medium, complex), we found the corresponding $p < 0.05$. The results have been furnished in Table 1.

7 DISCUSSION

7.1 Quantitative Result Interpretation

7.1.1 Hypothesis 1:

$H_0$: (Supported) We found a statistically significant difference in user preferences for the simultaneous controlling of the multiple degree of freedom i.e. while using 3 DOF or 5 DOF or AMB control. Though many participants agreed upon the fact that they are comfortable with a controller having 5-DOF, still they would like to have a better design of the controller which can ease their actions and provide the same functionalities.

"I would like to have a more controller where I can control the the three sliders could be replaced by control elements like scroll wheel or AMB control. Though many participants agreed upon the fact that they are comfortable with a controller having 5-DOF, still they would like to have a better design of the controller which can ease their actions and provide the same functionalities.

"Initially, I struggled to control the interface using both of my hands as I am not much familiar with these (ambidextrous) controllers. I might need more practice to control this interface (Sound stream)"

"It would be better if I could have a mouse like controller having three scroll wheels which I can use to change the tongue shape and move the mouse on a 2D space to change the tongue position"

Very few participants considered to have a controller with more than 5-DOF. "I would like to have a controller where I can control the source (glottal) frequency and gain with my legs so that I can have more sliders to control the tongue shape accurately with both hands"

To control the source frequency and gain, most participants preferred the mouse controller over the slider sensors ($M_{mouse} > M_{slider}$)

7.1.2 Hypothesis 2:

$H_1$: (Supported) We found a statistically significant difference ($p < 0.05$) in the user preferences while controlling multiple degree of freedom across all the three interfaces. As the VTDemo interface consists of multiple sliders only, the users could not control them simultaneously with a single mouse cursor. Though Pink trombone interface facilitates the simultaneous control of the source pitch and vocal tract shape change, the users did not find it intuitive at all.

7.1.3 Hypothesis 3:

$H_0$: (Supported) We did not find any statistically significant difference for creating different tongue shapes using all the three sliders. One possible reason could be the predefined structure of the vocal tract in VTDemo and Pink Trombone interfaces which is more intuitive for the users, as they provide a better mental model of which controlling parameters need to be changed to achieve the target. But this would be very difficult to generalize if we will increase the number of controlling points as user might take more time to adjust each controller to achieve a given shape.

7.2 Qualitative analysis and design issues

Based on the user feedbacks, we summarize the key points of comparison among the three interfaces as well as propose a few possible improvements of the design aspects of our interface.

The users found the Pink Trombone very engaging as it provides a better visual representation along with interesting sound variations. According to some of them, the interface is very clear and reactive and it simulates a real-tongue. A direct manipulation of the tongue for position control is convenient to use for educational purpose. Despite repeated questions being asked to the users around the ‘control’ aspect, the aesthetic aspect is seen to have a stable impact in their minds, which makes them inclined towards the Pink Trombone as far the convenience of control is concerned. However, users having previous background knowledge in vocal tract articulation find the changing of shape of tongue highly unacceptable in this interface. They provided some deeper insight that, despite looking like a real tongue, it somewhat fails to provide enough flexibility that the tongue is supposed to give. The tongue shapes and trajectories are preprogrammed, in a way that it can take limited number of shapes.

The VTDemo comparatively received low feedback, in terms of control and design. All the users found it hard to visualize the mapping between their slider control action and the final shape change. The lack of a direct control makes it confusing for them to interpret the effects of slider variations. The users also remarked that controlling one feature at a time makes it more unrealistic and it needs more coordination between the sliders to make it more interactive.

The Sound Stream gives a better physical interpretation of slider-based control to the users. Mostly, they seem to be contented with the ‘shape control’ side of the interface, bringing about various shapes with finger movements, which is evident from their comparative ratings as shown in Fig 10. However, according to some, the sliders could be replaced by control elements like scroll wheel or elements fitting into hand or fingers, that would increase the convenience of control, serving the same function. They find position control of the tongue relatively more difficult. This is because of friction arising from the paper base, on which the spring element lies, as well as the friction between the movable control block and the control space. So, it is necessary to somehow reduce the frictions...
of this prototype to help the platform glide more smoothly, that will increase the convenience of the proposed simultaneous control. However, the interface received an overall satisfactory response, as shown in Fig 11, on the basis of simultaneous articulatory and acoustic parameters, which was one of the targets of the work.

Users also felt that the width of the movable control block could be decreased and that of the control space could be increased to help them have a better grip and move it more freely. Based on their feedbacks, we have got a remarkable idea of replacing the square shaped movable block with a mouse-like structure, having three scroll wheels, i.e., a mouse with left and right click buttons replaced with scroll wheels. Obviously, it requires help from manufacturing experts and demands re-thinking of control idea, as the new arrangement would need motors and actuators to vary the tongue movement based on control command.

The users are more or less satisfied with the relative positioning and arrangement of the control blocks. Some suggested that bringing the left-hand and right-hand controls on the same horizontal axis could improve the convenience of control. While most users are satisfied with mouse based control, a few others prefer slider based control of gain and frequency. They justify that the mouse movement is highly sensitive and it is difficult to achieve precise control, for fine tuning of gain and frequency. Besides, they believe, there is more space for control of more acoustic parameters with their non-dominant hand, and hence one or two more sliders could fit in, which is not possible with 2D mouse-based control, which can only afford upto two degree-of-freedom control.

The availability of a five degree-of-freedom control arrangement is advantageous to the users, as evident from Fig 12, because it utilizes the functionality of the hand and fingers. However, we are not accustomed to do such rigorous tasks in our daily lives, that utilize extensive finger and elbow movements simultaneously. Thus, even though the users have the option of simultaneous changing of parameters, they would like to stick to two or three controls at a time. To summarize, the users realize that this arrangement gives them an opportunity to control more parameters at the same time, but increase in number of control also implies increase in training required to precisely control tongue movement. So, eventually, the applicability of these interfaces depend upon the user-types and the context of use. For new users or for learning purpose, users would like to use lesser control, however, for research purposes, they opine that the design nicely leverages the DOFs of control of both hands. The users also suggest to improve the aesthetic aspects of design so that it helps them to create a better mental model, which in turn will help the learnability and convenience of control.

In this prototype, we used a simple image processing algorithm to detect the structures and compute distances between them, for avoiding computational complexities and hence used black colour for the structures. The colour and texture of the background that represents the oral tract, the upper and lower boundaries representing the palate (hard and soft) and the lingual base could be changed to make it look more realistic. Besides, drawing artificial lips and throat and other mid sagittal structures would give the users a better understanding. In that case, more robust image processing algorithm needs to be used, that will require more computational time. So there is a trade-off between these two factors.

8 CONCLUSION AND FUTURE DIRECTION

In this paper, we explored the effect of simultaneous multiple degree-of-freedom control of tongue-like structure in articulatory speech synthesis. We assessed our interface with respect to Pink Trombone and VTDemo based on the proposed hypotheses. The users unanimously concluded that Sound Stream being a physical interface is much more intuitive and helps the user to feel the interaction better than the other two. Many users also opined that Sound Stream is better in terms of simultaneous controlling aspects which lies in the premise of our research question. However, there are still a few design issues in the prototype, like the dimensions of the control blocks, the friction between the spring element and the articulatory space or between the movable platform and the cardboard base, etc. So, we conclude that our prototype needs more robust design to make it more user-convenient and easy-to-use for practical controlling aspects. Nevertheless, conceptually, it can be seen to perform the targeted tasks better than the available interfaces in terms of controllability, expressibility and learnability and hence turns out to be the preferred interface, with respect to Pink trombone and VTDemo for simultaneously controlling a multiple DOF system, like the vocal tract.

The possible areas of improvement are the sound synthesis engine and some aspects of convenience and controllability. Firstly, we plan to implement Finite Difference Time Domain (FDTD) [8] based synthesis engine through graphics pipeline to leverage the high computational capabilities of a GPU as numerical integration of a second order PDE is necessarily a high computational problem. This makes the vocal sounds more natural sounding and intelligible without trading-off the execution time. The underlying reason for FDTD engine is, we are implementing a two dimensional vocal tract model, it would be better if we could somehow use a 2D synthesis engine rather than following the 1D area function based computations. Secondly, we also want to enhance the convenience and control aspects, as some of the users faced some particular difficulties to handle the Sound Stream interface. So, we would look into the design aspects and try to replace and rearrange the control elements based on the valuable suggestions received from the users. Furthermore, we would like to decrease the latency between mechanical control of the interface and the synthesis engine.

REFERENCES