Computers, Materials & Continua
DOI:10.32604/cmc.2022.022857
Article

Tech Science Press

A Real-Time Oral Cavity Gesture Based Words Synthesizer Using Sensors

Palli Padmini[1], C. Paramasivam[1], G. Jyothish Lal[2], Sadeen Alharbi[3,*] and Kaustav Bhowmick[4]

[1]Department of Electronics & Communication Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India
[2]Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India
[3]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[4]Department of Electronics and Communication Engineering, PES University, Bengaluru, India
*Corresponding Author: Sadeen Alharbi. Email: sadalharbi@ksu.edu.sa
Received: 20 August 2021; Accepted: 21 October 2021

Abstract: The present system experimentally demonstrates a synthesis of syllables and words from tongue manoeuvers in multiple languages, captured by four oral sensors only. For an experimental demonstration of the system used in the oral cavity, a prototype tooth model was used. Based on the principle developed in a previous publication by the author(s), the proposed system has been implemented using the oral cavity (tongue, teeth, and lips) features alone, without the glottis and the larynx. The positions of the sensors in the proposed system were optimized based on articulatory (oral cavity) gestures estimated by simulating the mechanism of human speech. The system has been tested for all English alphabets and several words with sensor-based input along with an experimental demonstration of the developed algorithm, with limit switches, potentiometer, and flex sensors emulating the tongue in an artificial oral cavity. The system produces the sounds of vowels, consonants, and words in English, along with the pronunciation of meanings of their translations in four major Indian languages, all from oral cavity mapping. The experimental setup also caters to gender mapping of voice. The sound produced from the hardware has been validated by a perceptual test to verify the gender and word of the speech sample by listeners, with $\sim$ 98% and $\sim$ 95% accuracy, respectively. Such a model may be useful to interpret speech for those who are speech-disabled because of accidents, neuron disorder, spinal cord injury, or larynx disorder.

Keywords: English vowels and consonants; oral cavity; proposed system; sensors; speech-disabled; speech production; vocal tract model

1 Introduction

Communication is essential in modern society, in every environment or social aspect of people's life, private or public. Statistics show that there are 400 million disabled people in the developing

world [1]. Approximately, more than nine million people in the world have voice and speech disorders [2–4]. Speech impediments are conditions in which normal speech is interrupted due to vocal cord paralysis, vocal cord damage, accidents, brain damage, laryngeal disease [5–7], laryngeal disorders [8], dysarthria, or cerebral palsy [9], neuron disorders, old age, oral cancer, muscle weakness, and respiratory weakness [10], etc.

Assistive technologies (AT) and speech synthesis systems that help the severely disabled to communicate their intentions to others and effectively control their environments, enable them to pursue self-care, educational, vocational, and recreational activities, etc. The present work aims towards the development of a mechanism to benefit patients with speech disorders based on oral cavity maneuvering only, without glottal and laryngeal intervention. Presently, there are different techniques to synthesize speech for the speech-disabled available in the commercial market, each with its own merits and demerits. A comprehensive summary of the state-of-the-art about these different techniques of speech synthesis system is listed in Tab. 1.

Table 1: Description of speech synthesis techniques

Techniques	Merits	Database depends on	Number of sensors/ electrodes used	For speech synthesis	Merits/Demerits
Electrolarynx [11]	Hand free device	—	—	Yes	It cannot synthesize the speech for the patients who have an oral cavity problem or spinal injury or neuron disorder.
Sign language to speech [12]	Portable, reduce costs and improve communication for deaf	Hand gestures	5	Yes	Need time to learn it first. It is not consistent for all languages. Symbols may be confusing. It can communicate only simple ideas

(Continued)

Table 1: Continued

Techniques	Merits	Database depends on	Number of sensors/ electrodes used	For speech synthesis	Merits/Demerits
Silent sound technology [13]	Allows speech communication without the use of sound produced by vocalizations.	Lip movements	4–16	Yes	In real-time, placing the sensors or electrodes on the face, the oral cavity is very difficult and is not easily possible in real-time
Articulatory speech synthesizer [14]	Mapping articulatory gestures. will give natural speech production as accurately as possible	Visual features	8–11	Yes	For laryngectomized patients, mapping articulatory gestures like velum are difficult, and real-time audiovisual information of speech
Vocal cord detection [15]	It detects the vocalizations which can control or access AAC devices.	Vocalizations	1	No	Very complex in holding the neckband for the whole day. Power management is high for the device to operate the vocal vibration switch. The vocal cord development switch is about $306.

(Continued)

Table 1: Continued

Techniques	Merits	Database depends on	Number of sensors/ electrodes used	For speech synthesis	Merits/Demerits
Brain implant [16]	Giving light to blind and paralyzed patients full mental control of limbs. Rescue missions.	Brain signals	More than 3000	Yes	Expensive. Risky in surgery. Wireless was not implemented yet. Difficulty in adaptation and learning
TALK [17]	Users can communicate repetitive phrases by just dictating a letter or two.	Exhales and inhales	1	Yes	Understanding and using the Morse code for each of the alphabets is complex. Taking short and long exhales is difficult while the people having asthma, allergies, sinusitis, cough, chest congestion, runny nose, or pulmonary disease
Tongue drive system [18,19]	Easily access a computer or control a motorized wheelchair, robotic arm, phone, TV	Tongue movements	4–6	No	Difficult to hold a magnetic tracer sensor on the tongue. It is not useful for speech production

(Continued)

Table 1: Continued

Techniques	Merits	Database depends on	Number of sensors/ electrodes used	For speech synthesis	Merits/Demerits
Our Proposed System	Easily capture the oral cavity movements using sensors	Tongue, teeth, and lip movements	4	Yes	It can be used for different degrees of speech-disabled, multiple disability children, an Easy way to produce speech only using tongue movements. In real-time can use salvia proof touch sensors

The summary hence drawn, is that in the aforesaid literature, speech synthesis systems have mostly used electrodes or sensors or visual features information from lips or hands, that are captured for speech production. The tongue drive system was explored in recognizing the oral cavity activities and to control the remote devices, but although promising, it has not been explored for speech production. Based on tongue gestures, author(s) have previously estimated optimized formants of the oral cavity for the English alphabet, to prove that the tongue plays a key role in speech production [20]. From the basics of phonetics, and articulatory system-based method was proposed by author(s) initially (including tongue, teeth, and lips) with listed accurate gestures for each English alphabet, and was tested for English alphabets using LabView [21]. However, an initial practical demonstration was not published by the author(s) until later [20], although yet not for complete words.

The main objective of the proposed work is to synthesize sound/words based on the movements in the oral cavity (tongue, lips, and teeth) during the production of human speech, especially the articulation of English vowels, consonants, and sequence of letters i.e., words. The tongue, which has only been clinically characterized thus far, has been considered as the main player in English vowel and consonant production. Refrences [20,21], by the author(s), previously, leading to the present work. In this paper, a tongue-based system has been conceptualized for producing human speech to create a demonstrable device to produce up to words. Thus, the present solution demonstrated can potentially only help people with speech disabilities in social life, and also aid in their healthcare, professional life, family life, etc.

The main contributions of the work are as follows:

- The oral cavity gestures need to be identified for each vowel and consonant during the production of human speech.

- Optimization of the number and position of the main sensors needed to implement tongue-driven speech synthesis. The optimized setup is implemented via an in-house built device and the performance is experimentally demonstrated.
- Implementation of the concept of the proposed system in hardware for use in real-time applications to allow speech-disabled people to produce words in English and major Indian languages.

The remainder of the paper is organized as follows: Section 2 explains the system of human speech production, followed by the proposed system in Section 3. Section 4 describes the setup of the experimental hardware, followed by a conclusion and expectations for future research in Section 5.

2 Oral Cavity Gestures Identification Based on Human Speech Production System

In the speech production model, the air is forced from the lungs by the contraction of muscles around the lung cavity. The pitch of the sound is caused by the vibration of cords. The excitation of the vocal tract results from periodic puffs of air. The vocal tract (between the vocal cords and lips) acts as a resonator that spectrally shapes the periodic input that can produce the respective sound from lips [10].

The parts of the vocal tract that can be used to form sounds or during speech production are called articulators including the glottis, larynx, jaw, tongue, and lips, teeth. The articulatory gestures of speech production in existing speech production models involve the glottis, larynx, tongue, lips, teeth, and palate for producing specific sounds. The first step was to understand and analyze the existing mechanism to produce speech in the vocal tract by using VocalTractLab 2.3 software (VTL 2.3) [22]. It identifies the articulators, which play an important role in improving acoustic simulation in VTL 2.3. This model represents the surfaces of the articulators and the vocal tract walls. The shape and/or position of the articulators is defined by vocal tract parameters like glottis, larynx, jaw, tongue, teeth, and lips.

In general, the shapes of the vocal tract during articulation of a few English vowels and consonants in the existing model using VTL 2.3 software are shown in Fig. 1. This makes it easy to visually compare the vocal tract model shapes for the pronunciation of English syllables, especially if they are displayed as 2D or 3D contour images. The tongue, teeth parameters define tongue height (h) and tongue frontness (l) which specify the tongue shape for each English alphabetic sound [20]. The degrees of lip rounding and velum is the soft tissue constituting the back of the roof of the mouth called soft palate, which is specified by the parameters lip rounding and velum position. When a parameter (tongue, teeth, and lips) is changed, the corresponding changes in the vocal tract shape, observed in Fig. 1, for a few English vowels and consonants.

Again, vocal tract (VT) animation with the glottal area, larynx height, jaw height (which resembles tongue displacement), tongue position, tongue tip/apex, and lips from vocal tract acoustics demonstrator (VTDemo) software [23] is shown in Fig. 2, for the articulatory synthesizer.

VTDemo shows the vocal tract positions during sound synthesis. When articulatory parameters change, it changes the sound that is heard. By varying the height of the jaw or the displacement of the tongue body, tongue position, tongue tip, and lips in the VT Demo articulatory synthesizer software, researchers can observe the production of different vowel sounds [23]. The shapes of articulatory gestures during the production of English vowels and consonants using VT Demo are shown in Fig. 2, which helps to identify the oral cavity gestures of each alphabet.

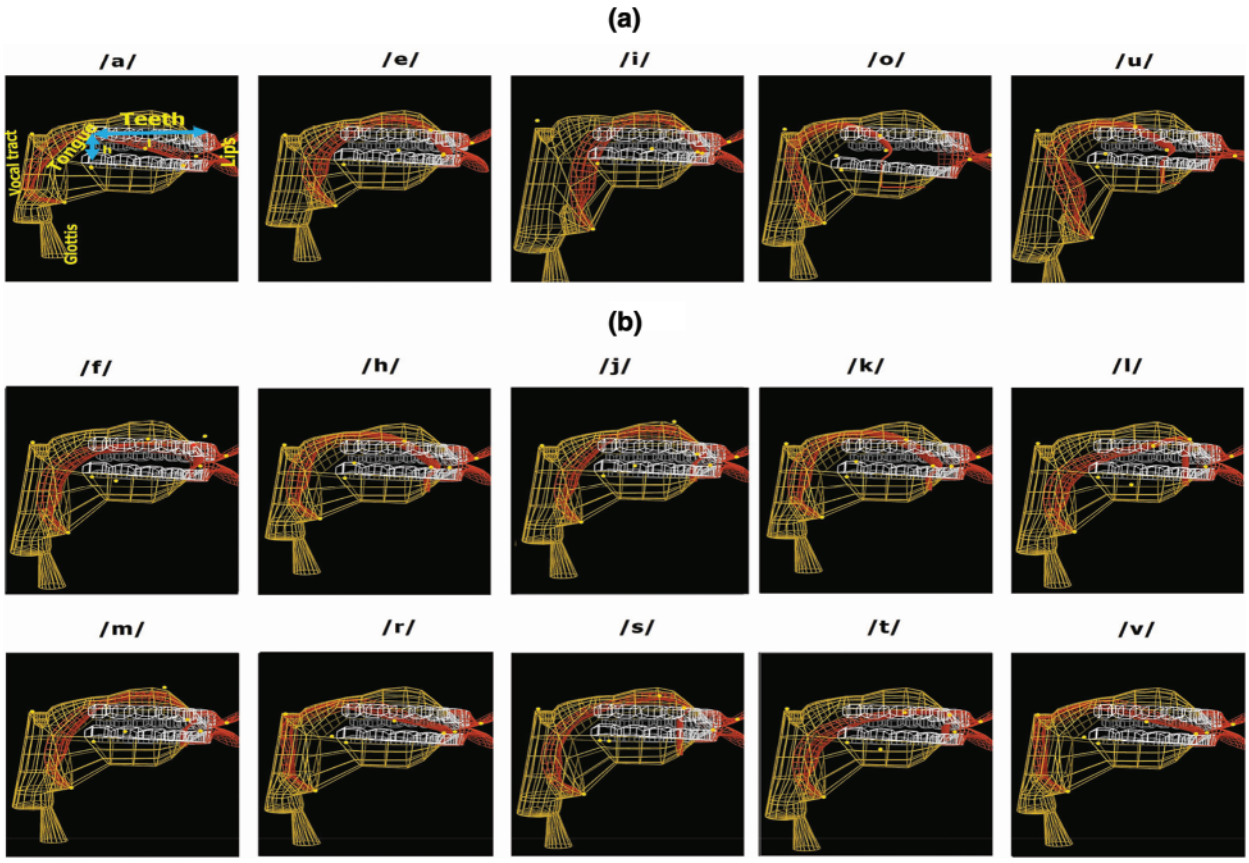

Figure 1: 3D representation of the vocal tract model while pronouncing English (a) Vowels and (b) Consonants using VTL 2.3 software [22]

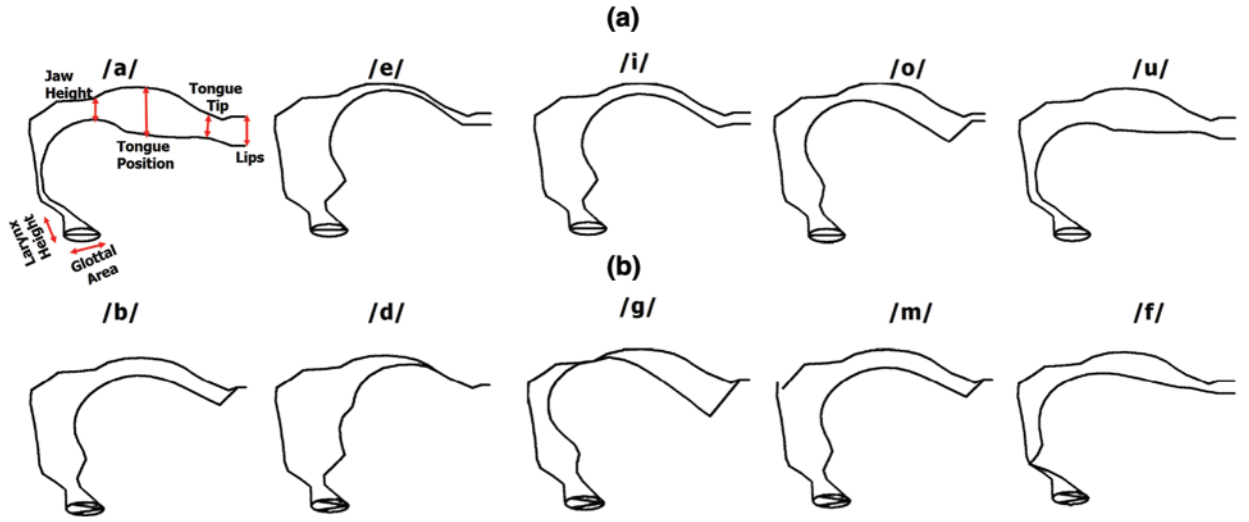

Figure 2: Oral cavity gestures while pronouncing a few English (a) Vowels and (b) Consonants

The observation and analysis of the existing vocal tract speech production systems using VTL and VTDemo as shown in Figs. 1 and 2, for both vowels and consonants help us to estimate the correct locations/gestures to capture the parameters of oral cavity like tongue, teeth, and lips position data for speech production for each English alphabet without the use of glottis and larynx. Thus, the focus of the study is to use only oral cavity movements (tongue, teeth, and lips, jaw), with the tongue being most important [20] to produce sound. This requires building the proposed sensor-based system for speech production for the speech-disabled using only oral cavity gestures. It is described in the following sections.

3 Proposed System to Synthesize Speech Using Matlab GUI

From the literature review, we found that there are not many studies that focus only on the oral cavity articulatory gestures for speech production, except in author(s) work [21]. Thus, the proposed system is built by using the human speech production mechanism by using only the gestures of the oral cavity like jaw height, lips, tongue body displacement, the position of the tongue, and the tip of the tongue, without using the glottis and larynx. The present work concentrates mainly on the functionality of the lips and tongue (in the oral cavity) in the production of English vowels and consonants. The production of each sound depends on the degree of lips i.e., lip movements, tongue tip, and tongue body displacement, and tongue teeth positions.

The front and side views of the oral cavity are shown in Figs. 3a and 3b, respectively. We estimate the four sensor positions of the proposed system for speech production are highlighted by red dots is shown in Fig. 3b.

Moreover, a set of gestural descriptors is used to represent the contrastive ranges of gestural parameter values discretely [24]. These descriptors point is a set of articulators involved in a given gesture and the numerical values of the dynamic parameters, which characterize the gestures. Every gesture can be specified by distinct descriptor values for the degree of constriction. Fig. 4, shows the gestural dimensions and their descriptors in detail, including a comparison with current proposals of feature geometry [24].

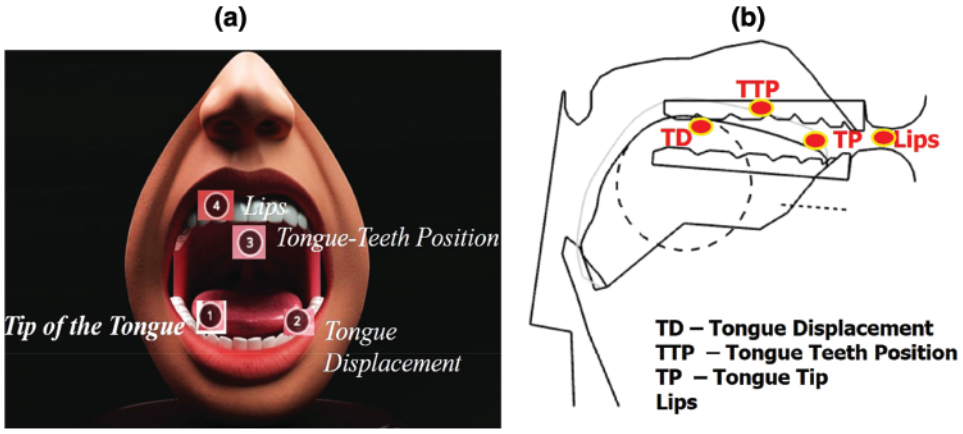

Figure 3: Estimated four sensors positions in the oral cavity (a) Front view (b) Side view

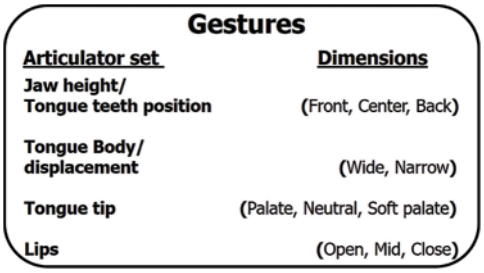

Figure 4: Inventory of articulator sets and associated parameter [24]

Oral gestures involve pairs of tract variables that specify the degree of constriction of the tongue. For simplicity, we refer to the sets of articulators involved in oral gestures, viz., the lips, tongue tip (TP), tongue teeth position (TTP), and jaw height (JD) for gestures involving constriction of the tongue body. These gestures will help to differentiate each English alphabet easily. Individual gestures of each sensor position with different articulatory gestures are successively shown in Fig. 5.

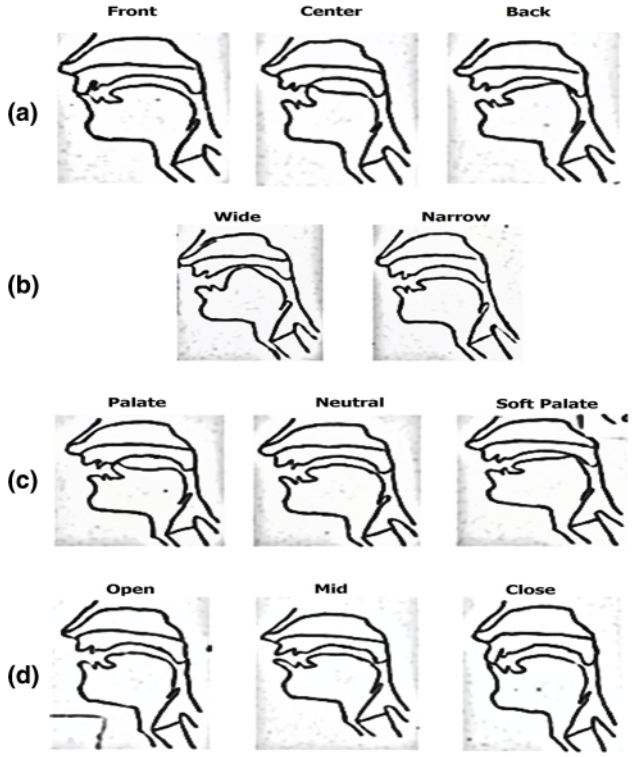

Figure 5: Schematic representation of (a) Tongue-teeth position (b) Tongue displacement (c) Tongue tip (d) Lips while producing English vowels and consonants [10]

Data from all four sensor positions that capture the gestures of the tongue, teeth, and lips in the oral cavity along with the side and front view of the oral cavity during the production of each English vowel is shown in Fig. 6. For example, to produce the sound '/a/,' the tongue tip is neutral, tongue displacement is narrow, lips are open and the tongue-teeth position is back as highlighted by red dots.

Generalized sensor-based inputs (i.e., jaw height, tongue body, tongue tip, and lips) for a few English vowels and consonants are given in Fig. 6.

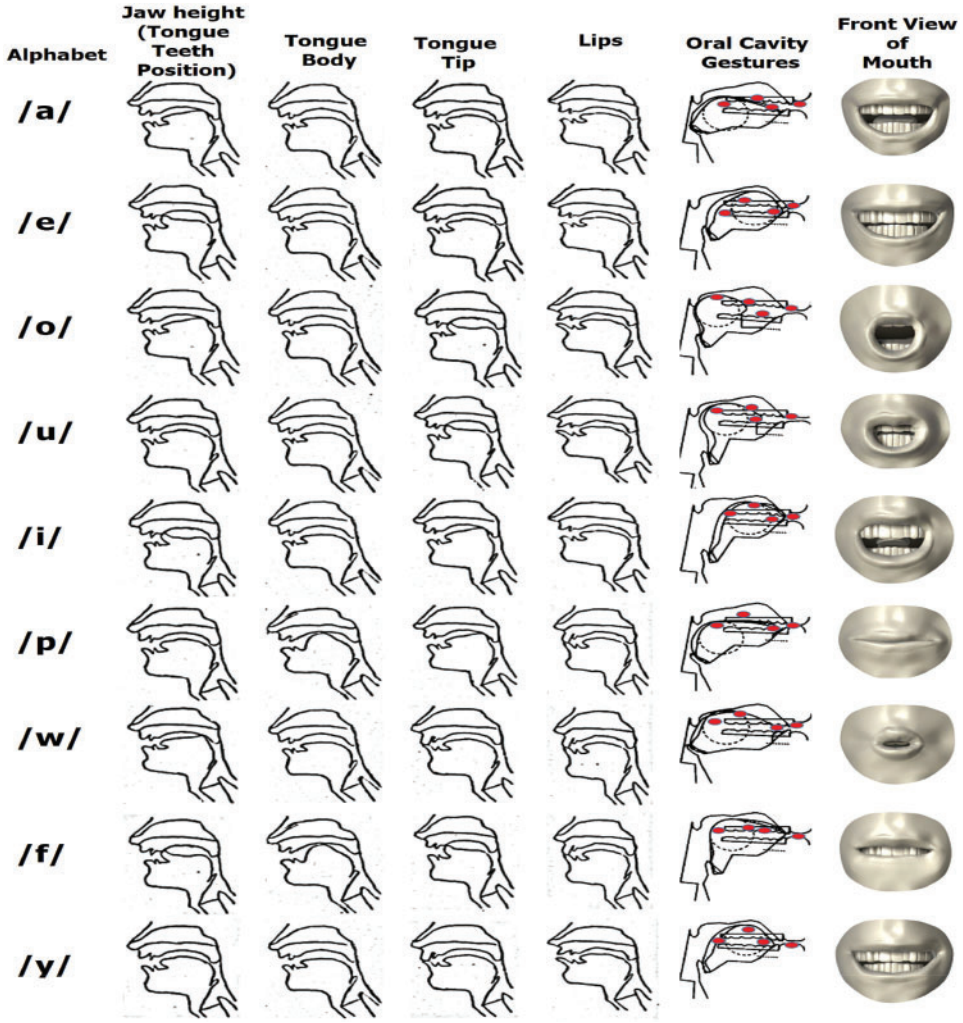

Figure 6: Four sensor positions along with a front view of the oral cavity gestures for a few English vowels and consonants

Likewise, we can produce the sounds of either English vowels or consonants using the captured four sensor positions of oral cavity gestures. Different combinations of four articulatory gestures produce different English sounds. Irrespective of these combinations of four articulatory gesture inputs, no sound will be produced.

The articulatory gestures through sensor input are monitored and captured continuously to synthesize the sequence of vowels, consonants i.e., words. The gestural score for the utterance of each word is based on articulatory phonology, as shown in Fig. 7 [25]. The rows correspond to distinct organs (JH = "Jaw Height," TD = "Tongue Body," TP = "Tongue Tip," Lips). The labels in the boxes stand for the gesture's goal specification for that organ. For example, alveolar stands for a tongue tip constriction from the horizontal. Each syllable connects critically coordinated gestures, or phased, for one another that represent greater bonding strengths between coordinated gestures [25]. The gestural

score for uttering a few words, with boxes and tract variable motions as generated by the computational model (coordinate pairs of gestures), is shown in Fig. 7.

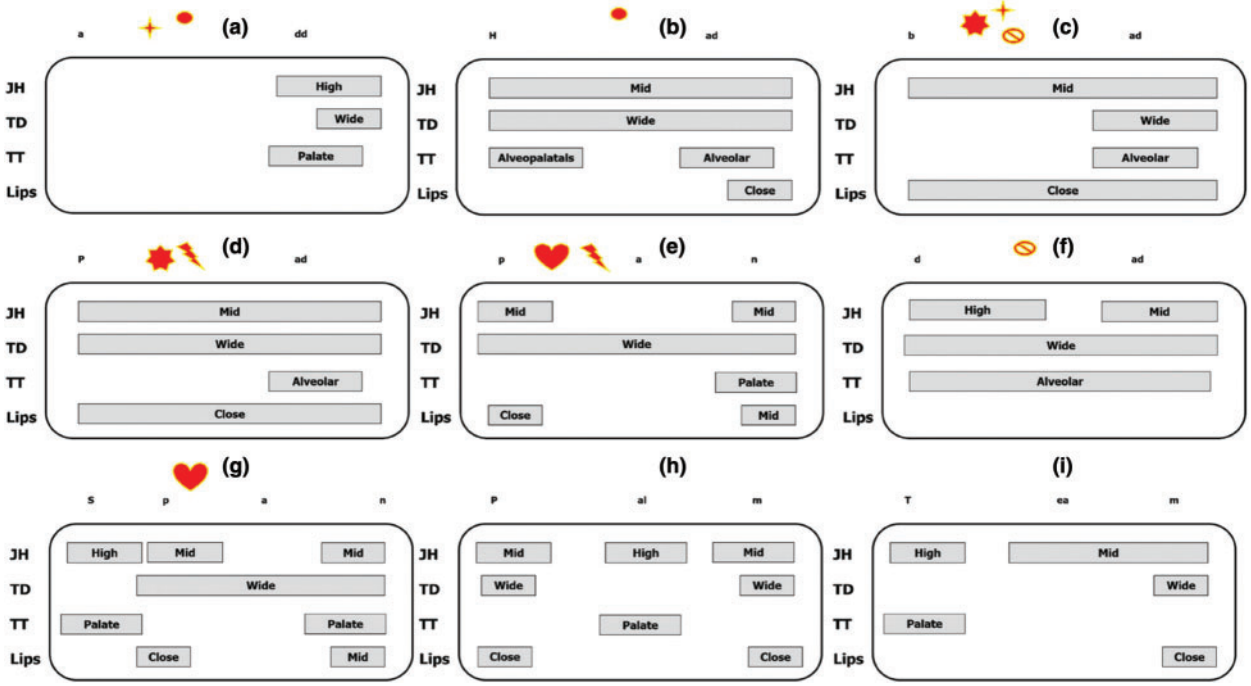

Figure 7: Schematic gestural scores. (a) "add" (b) "had" (c) "bad" (d) "pad" (e) "pan" (f) "dad" (g) "span" (h) "palm" (i) "team"

Fig. 7 can be substantiated by displaying the gesture scores while reciting each particular word. Philological items contrast gesturally, by verifying each gesture is present or absent (e.g., "add" *vs.* "had," Figs. 7a,7b; "add" *vs.* "bad," Figs. 7a,7c; "bad" *vs.* "pad," Figs. 7c,7d; "pad" *vs.* "pan," Figs. 7d,7e; "pan" *vs.* "span," Figs. 7e,7g). Those combinations are pointed and highlighted with different red shapes. We assume that in speech mode "had" and "bad" would typically be considered to differ from "add" by the presence of a segment, while "bad" and "pad," "pad" and "pan," would differ only in a single feature, voicing or nasality, respectively. Another kind of contrast is that in which gestures differ in their assembly, i.e., by involving different sets of articulators and tract variables, such as lip closure *vs.* tongue tip closure (e.g., "bad" *vs.* "dad," Figs. 7c,7f). All these differences are categorically distinct.

The response of the complete proposed system for speech production was validated by creating a GUI using Matlab. In our proposed system, four-sensor data inputs i.e., tongue body displacement or jaw height, tongue position and tongue tip, lips positions were used for the speech synthesis system. We have made a list of the position of each sensor like as shown in Fig. 6, for each English alphabet. If the combination of four sensors positions matches with the pre-defined look-up-table as in Fig. 6, it has to display the respective letter in the GUI and produce the same sound. The four sensor inputs are selected manually using the drop-down arrow, GUI for speech production is built for English vowels and consonants using Matlab, as shown in Fig. 8. The steps to build GUI using Matlab is as follows,

1. First select the sensor inputs manually using the drop-down button.
2. The sensor inputs are checked with a pre-defined table, based on match, the respective letter displayed in the text box.
3. we use text to speech conversion in Matlab, to produce the speech sound.
4. For the sound produced, the pitch will be calculated and display the sound waveform on the screen.

Thus, the GUI shows four sensor inputs, and the respective letters of the alphabet are displayed on the text box, speech signal waveform, pitch of the sound, and the equivalent sound. For example, if our sensor inputs are open, front, narrow, and neutral are selected, according to the predefined table, these combinations should produce/c/ sound as is shown in Fig. 8b.

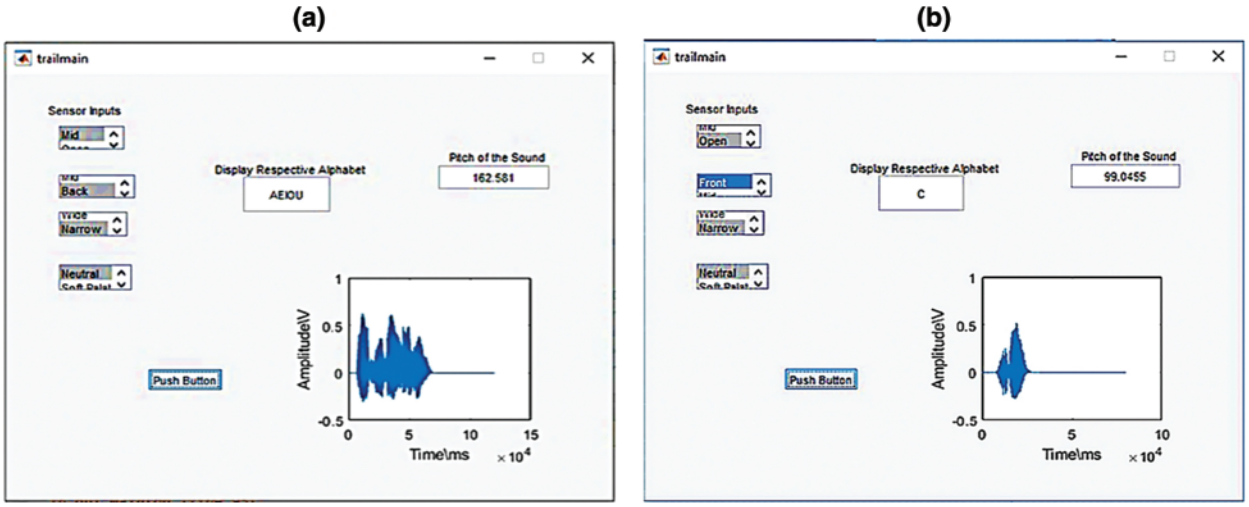

Figure 8: GUI of the proposed system using Matlab for speech synthesis based on the sensor–input of (a) Sequence of vowels (AEIOU) (b) Consonant Production

The sensor-based input continuously monitors the position of tongue, tongue-teeth, and lips during the articulation of a sequence of vowels and consonants using text to speech synthesis (TTS), whose GUI is shown in Figs. 8a and 8b, respectively.

In our proposed system, speech is synthesized by capturing the different articulatory gestures of the oral cavity (tongue, lips, teeth) based on sensor-based input data. It brings out a hardware design with each of the roles of articulators for speech production using appropriate sensors and electronic devices. Furthermore, we implemented the sensor-based hardware experimental setup version of the proposed system. The hardware setup is described and discussed briefly in the following section.

4 Real-Time Speech Production System for Speech-Disabled Using Tongue Movements

A speech-impaired person can be able to produce/pronounce a letter/word distinctly using tongue movements. In our proposed hardware system, speech is produced by using the four sensors, which are placed inside a prototype tooth model, to capture tongue movements. We observed the bent and rolling of the tongue in numerous degrees and their subsequent touch is recorded. The estimated lookup table (LUT) is to be coded in the microcontroller's memory. The produced letter is played over a speaker using parallel communication with a micro-SD card. This section describes the components of the hardware and the flow of the algorithm and features of the proposed hardware system. It also analyzes the output of the proposed hardware system, which is validated by a perception test.

4.1 Proposed Hardware Experimental Setup

The proposed system was initially limited to the pronunciation of the English alphabet and words by using a prototype model of teeth, perceived as a biological system, and extended to other languages using a language translator. In our proposed system, speech production happens only from movements of the tongue inside the oral cavity. Based on knowledge of existing speech production systems, we concluded that four-sensor positions are required for speech production using oral cavity movements, as discussed in Section 3.

The hardware prototype is improvised to sense jaw movements and tongue flexibility. In the proposed system, a set-up to analyze the speech sound from a laptop and LCD screen is shown in Fig. 9a. To capture the tongue and jaw movements (including tongue teeth position and lips), we use two limit switches, a potentiometer, and a flex sensor, as shown in Fig. 9b. One can track the position of the oral cavity movements during articulation displays and produce the respective sounds in an LCD of speech assistant device and through an electric speaker.

The components required to set up the mini-prototype to synthesize words with a rechargeable battery have Arduino UNO, Bluetooth module, flex sensor, potentiometer, connecting wires, bread-board, and a 9v DC battery with connectors, as shown in Fig. 9b.

Two limit switches were placed in the upper and lower palates to indicate whether the tongue touches the palate. The potentiometer is placed on the tooth set to sense the jaw movement and lips positions, and the flex sensor is placed on the tongue to identify the twist and roll positions of the tongue, as shown in Fig. 9b.

It is assumed that the vocal tract is not present because the person has a larynx disorder or a voice box problem. The algorithm is designed to observe the tongue, teeth, and lips functions and capture the same with tongue movement, tongue touches, and jaw movement. A unique algorithm was created to pronounce every letter of the alphabet distinctly. A LUT is formed using the gestures made during the articulation of each vowel or consonant for digital data acquisition. Optimized coding is incorporated for every sensor calibration on sensible data to run the algorithm. The prototype of the speech assistant system has been designed to fit the small-scale model of human dummy teeth with minimal circuitry. The flow diagram for the real-time speech production algorithm is shown in Fig. 10.

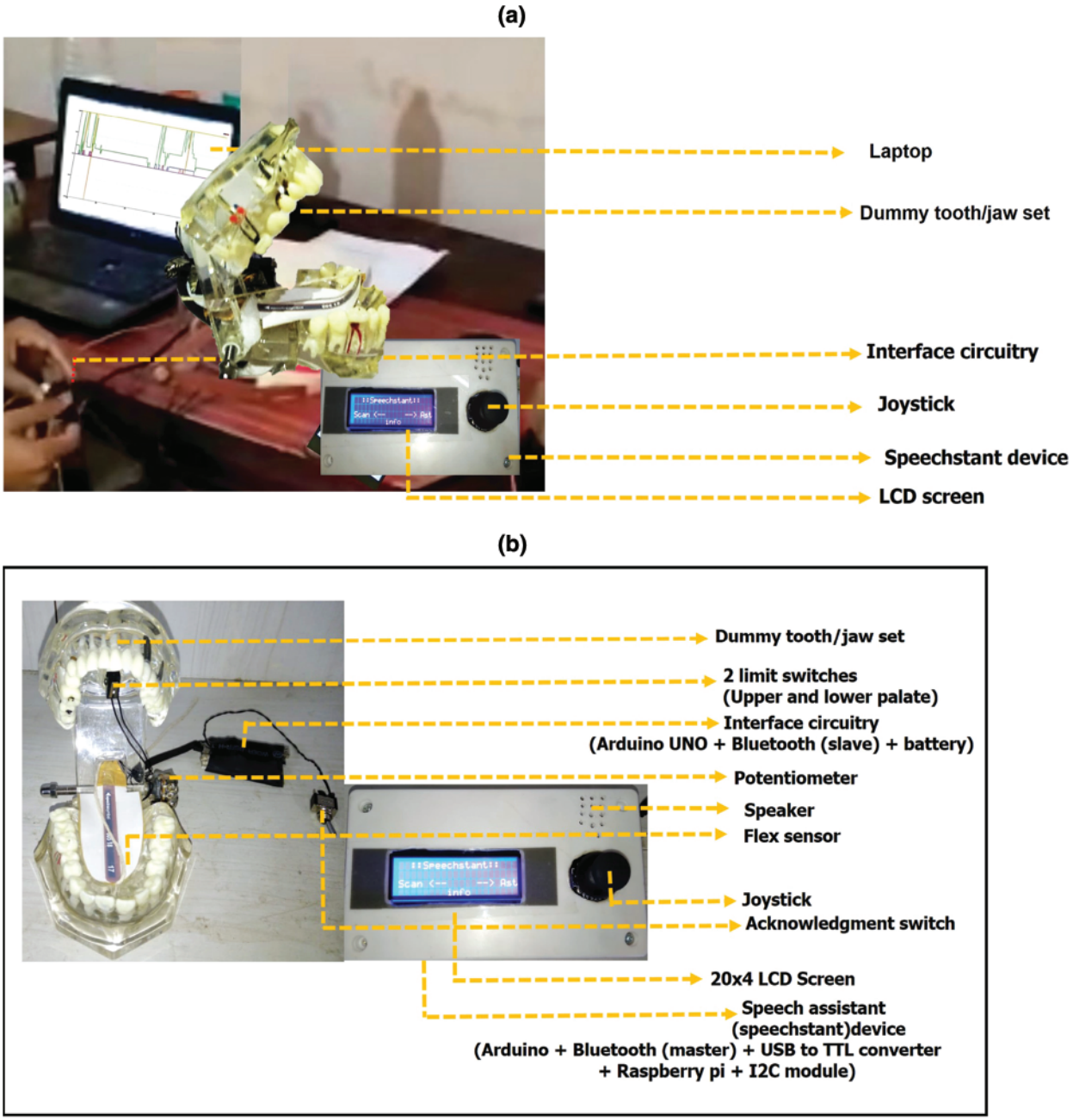

Figure 9: (a) The hardware experimental setup (b) Hardware components of the proposed system for sound/word production

There is two independent (dummy tooth/jaw set and Speech assistant device) and one interdependent (interface circuitry) aspect of the hardware. They are treated in what follows.

- Dummy tooth/jaw set: This is used by dentists to demonstrate how teeth, cavities, and gums, interrelate. It is used as a dummy human tooth/jaw set. In this work, the tooth set is used as a medium to fix the sensors and acquire data. We have placed different sensors in different

positions to capture the relevant oral cavity gestures during each alphabet production but considered sensors are in optimal positions which gives good information of gestures in identifying the sound production which improves the accuracy/performance of the system. Thus, during trials on experiments, we consider four sensors (S1-S4) affixed to the dummy tooth/jaw set to acquire data/values that are sufficient/required to produce a particular letter or word.

- Interface circuitry: This circuitry enables the communication between the dummy tooth set and the speech assistant device. The input sensor values from the dummy tooth set are sent to the speech assistant device.
- Speech assistant device: This is the final part of the hardware and also the most important part. Its basic features are:
    - o It stores letters and words.
    - o It has a user-friendly graphical user interface (GUI).
    - o A control system allows users to choose words from lists and announce them to fellow listeners.
    - o With the use of a dummy tooth/jaw set, it can save the pronounced word for future use.
    - o Machine learning and natural language processing libraries were incorporated for the synthesis of words in any language that Google Translate can enlist.
    - o Offline and online modes are available for the user's benefit.

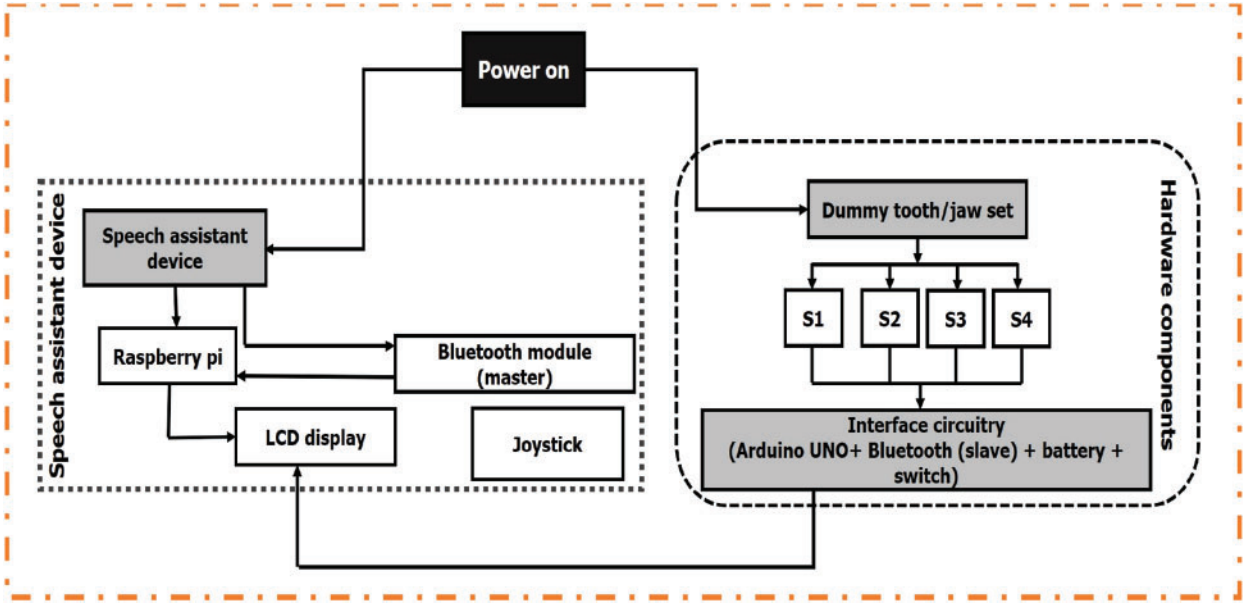

Figure 10: Flow diagram of the proposed system for sound production

The hardware components required for the experimental setup are enumerated as given in Tab. 2.

Table 2: The hardware components required for the experimental setup

Hardware component	Functional	Why
Dummy tooth/jaw set		
Limit switch	This is affixed at the upper and lower palate of the dummy tooth set to register touch that is induced by the tongue on either the upper or lower palate. The dimensions of the limit switch consider are 20 mm × 10 mm × 7 mm with 5A and 250 V switching current and voltage respectively [26].	It represents the tongue not touching or touching the palates (upper and lower), respectively. The reason any other thin piezoelectric sensor is not used is that they register values based on continuous pressure whereas the human tongue does not apply continuous and constant pressure to either of the palates, so such a sensor would provide false logic.
Flex sensor	It is used to capture tongue movements. It is set for analog values from 0 to 5. 0 means there is no tongue movement, and 5 means the tongue tip touched the upper palate for certain letters of the alphabet. The dimensions are 7.3 cm long and 6.3 mm wide [27].	Tongue movement involves bending and rolling, so the closest sensor that can replicate such movement is the flex sensor.
Potentiometer	The pronunciation of letters also requires upper and lower jaw movement. The value of the upper jaw is registered as (0–5). The lower jaw is almost always immovable, hence adding benefit to the user. The dimensions are 23 mm in diameter, ranges from 10–10 kΩ [28].	The movement of the tongue also is analog and with every varying movement, the effect of the pronunciation of any particular letter is different. The best way to capture the various jaw movements is by using a variable resistor that gives a variable output voltage

(Continued)

Table 2: Continued

Hardware component	Functional	Why
Tongue made of a flexible visiting card	The dummy tongue is modeled to replicate the movement of the human tongue as far as possible.	The proposed tongue is hand-made and is cut from material that is used for flexible visiting cards. The dummy tooth/jaw set provided by the user is too small, so using bare hands inside the model will be a challenge, so one can use a thread to let the tongue move. According to the movement, the flex sensor can register values.
Interface circuitry		
Arduino Nano V3.0	This is used to program the Bluetooth module HC05 for transmission of data acquired from the dummy tooth/jaw set by sensors.	It is very small and compact, and it has all the functionalities of Arduino UNO.
Bluetooth module-HC05 (Slave)	Communicate wirelessly from the dummy tooth set to the speech assistant device through this module. This is used as a slave to the Bluetooth module that is connected to Raspberry Pi.	The user will have to face difficulties in connecting the dummy tooth set to the speech assistant device by using wires. This provides better mobility and portability options for the user. The wireless, and Bluetooth is the best option.
3.7 V Lipo Battery	Since the battery is small, it can be easily integrated with the interface circuitry and wrapped in a heat sink pipe so that the circuit looks neat.	It is efficient and ideal for powering the Arduino Nano V3.0 and Bluetooth module for 30 min. For more usage, we need bulk batteries, which cannot be made part of the compact user-friendly wearable circuit.

(Continued)

Table 2: Continued

Hardware component	Functional	Why
Acknowledgment or Priority switch	This is used only to lock onto the desired sensor value among the running values that are displayed on the mobile screen. Then, we move to the next sensor and perform the same operation. When the values are matched from the lookup table (LUT), one can leave the priority switch, and the letter is stored. Then we can move on to store more letters and store them as complete words.	This is for the convenience of the user so that once a desired value from the sensor is attained, the priority switch can store the value and the user can see the stored value on the mobile screen. For the next sensor value, the user does not have to worry about the previously stored value.
Speech Assistant (speechstant) device		
Bluetooth module –(Master)	In this case, the Bluetooth module HC05 is used as a master. It connects with its slave pair (discussed in interface circuitry), which is connected to the dummy tooth/jaw set and receives the data from the sensors.	Since the speech assistant device is a wireless system that uses the Bluetooth module, the sensor data coming from the dummy tooth set can be received only by another Bluetooth module that is connected to the speech assistant device.
USB to TTL converter	This is used to collect data from the master Bluetooth module and send them to Raspberry Pi.	Bluetooth modules can only transmit and receive data. They do not store data. This device must store words, letters, and numbers; those data must be stored in the memory card of Raspberry Pi. Thus, it is essential to use this module to send the data from the Bluetooth module to Raspberry Pi.

(Continued)

Table 2: Continued

Hardware component	Functional	Why
Raspberry Pi	This is the heart of the project. It communicates with the dummy tooth/jaw set, and it enables the user to store letters or words and announce them. (a) A microcomputer can be connected to a monitor and then can be accessed as an Ubuntu operating system. It is a very robust and secured microcomputer. (b) It has a faster processing capability because of parallel processing.	Raspberry Pi can perform multiprocessing operations like: (a) Collecting data from a USB port via Bluetooth (b) Processing text to speech (c) Checking parallel Wi-Fi connectivity (d) Managing remaining peripheral interfaces like display, joystick, audio driver, file handling. Arduino does not do parallel processing. So, all these tasks would take much longer for Arduino to perform, and that is not acceptable.
Joystick	This is an added component that gives a user-friendly outlook to the speech assistant device. In this case, it gives the user the freedom to maneuver and select the options shown on the LCD. The buttons that are enabled are left, right, up, and down for navigation, and the center can be pressed for storing	It is easier and more comfortable for a user to use the joystick instead of pushing buttons.
Arduino Nano V3.0	This is used in the speech assistant device to convert analog signals from the joystick to digital signals for Raspberry Pi. The input of the Arduino Nano is an analog signal from the joystick. It converts the analog signal to a digital signal to enable the cursor movement that is observed on the LCD screen.	Without it, Raspberry Pi cannot receive the signals from the joystick.

(Continued)

Table 2: Continued

Hardware component	Functional	Why
I2C module for 16 × 2	This is used for communication between Raspberry Pi and 20 × 4 LCD using a 2-wire protocol, namely I2C.	The user must see the data in the LCD. The data from the dummy tooth set is acquired by the master Bluetooth module, and Raspberry Pi gets the data. This module communicates and displays the data on the 20 × 4 LCD. Without this, the LCD could not communicate with Raspberry Pi or display any results.
20 × 4 LCD	This is the display component of the speech assistant device. All graphical components of the project, including Scan, Assistant, and Info, can be seen in this display. Any sensor data acquired from the dummy tooth set can be seen in this display. This is used to search out a particular word or number or letter that the user has stored.	It is small and cost-effective. Also, it fits exactly inside the speech assistant device. Thus, it is portable. Since only four words can be seen at a particular time, it is easy for the speech-disabled person to use this display. Mobile screens could also be fit for the display, but more modules are required to establish communication between the dummy tooth/jaw set and Raspberry Pi.
Power source	This is mandatory for the speech assistant device to work. It can be connected by USB or be a power bank of 10,000 mAh.	It can be used to power the speech assistant device. When it is not required for this project, it can serve as an excellent power recharger tool for any electronic gadget.

Two limit switches, flex, and potentiometer sensors were placed in the oral cavity to capture the oral cavity gestures. Based on oral cavity gestures, we capture and obtain data by affixed sensors on a dummy tooth set. The oral cavity gestures at different positions during articulation of different sounds are shown in Fig. 11. The different oral cavity gestures of the proposed hardware system are shown in Fig. 11, as same as the oral cavity gestures discussed in Section 3 and shown in Fig. 6. The dummy tooth/jaw setup is activated to exhibit various gestures associated with different letters and words of the English language, a few of the gestures are shown in Fig. 11.

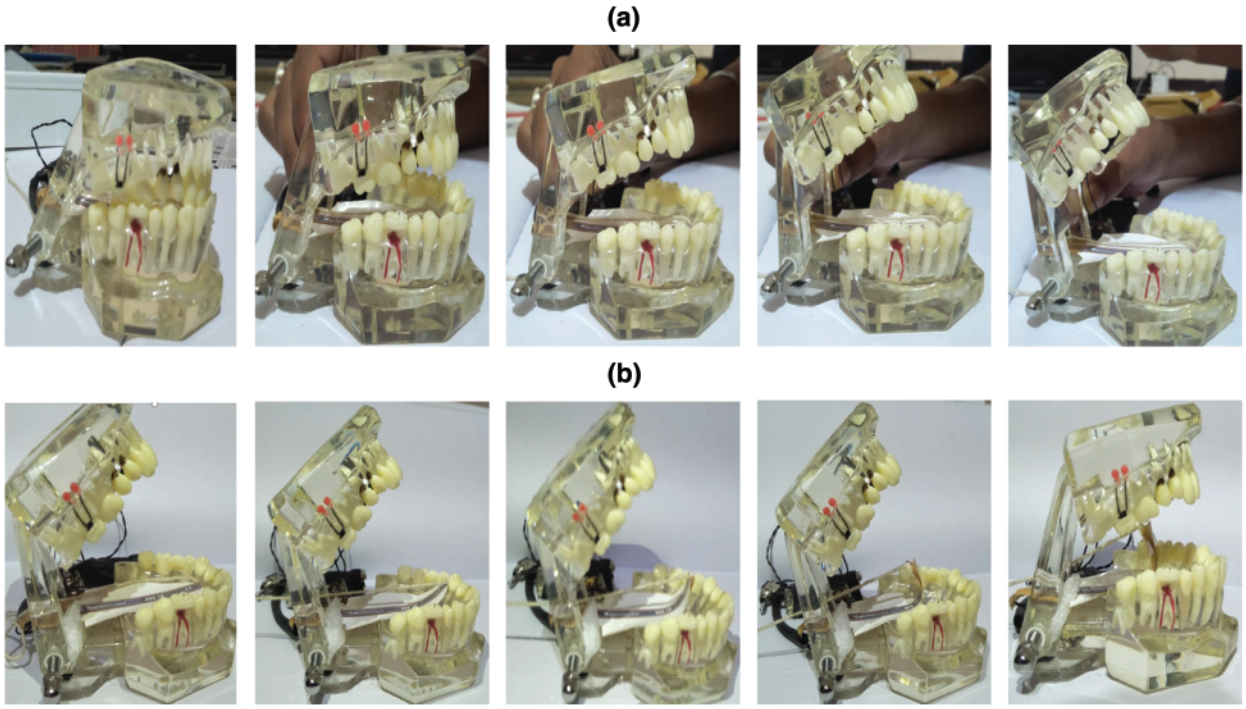

Figure 11: (a) Multiple degrees of Jaw gestures (b) Various tongue height and advancement gestures

The input sensor data values captured from the different oral cavity gestures were sent to the interface circuity and speech assistant device to pronounce a sound. The respective output speech would be heard from the speaker based on a match of the variability of four sensor values with LUT. Tab. 3 shows the LUT values in centimeters of four input sensors for a few English letters. LUT is used for interfacing the sensor outputs values in accomplishing speech synthesis of our proposed hardware setup.

Table 3: The LUT values of four sensors for a few English alphabets

Look-Up Table (LUT)															
	English letters														
	/a/	/e/	/i/	/o/	/u/	/j/	/w/	/p/	/b/	/t/	/d/	/k/	/g/	/m/	/n/
S1	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1
S2	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1
S3	3	2	2	4	5	4	4	3	3	4	5	5	3	2	3
S4	1	2	4	3	4	1	5	3	4	2	3	2	2	5	5
Tolerance/error we have considered as +/− 0.5															

The input sensor data was tested using a LUT (see Tab. 3). The LUT values of sensors 1 and 2 (S1, S2) define the position of the tongue whether it touches the upper palate and lower palate

or is not defined by 0 or 1. Sensors 3 and 4 (S3, S4) define the tongue teeth position and tongue tip using flex and potentiometer as shown in Tab. 3. As the sensor devices, we have considered are electrical devices, thus tolerance/error will be considered. We have considered the tolerance values are +/− 0.5 of sensor values, to get effective results. The purpose of these tolerance values will help us to get effective information from the sensor, even sometimes the electrical sensors/devices get affected by temperature/heating issues.

If input sensor data match with LUT, a respective letter will be displayed on the screen. If sensor data did not match with LUT, no letter was printed on the screen. The respective displayed letter was passed to the function called for audio play through serial-parallel interface communication with SD card (flash memory). The letter sounds, which had been stored, came through the electric speaker. The steps were repeated while taking the sensor data continuously to produce a letter or sequence of letters (words) and the relevant letter is printed on the LCD screen.

4.2 Features of the Proposed Hardware System

The proposed system is completely built keeping given the original idea of pronouncing alphabets and extrapolating it to synthesize words. The features of the experimental setup are enumerated as follows.

- Prototype of small-sized dummy teeth model: The speech assistant system is designed to fit the small-scale human teeth model with minimal circuitry. It has been designed to follow the same LUT. The model is now more realistic and easier to use.
- Interface of the dummy teeth model with the GUI-based speech assistant: The dummy teeth model interfaces with the Raspberry-Pi-enabled speech assistant system wirelessly. The data collected from the sensors can be sent via Bluetooth to the speechstant system, and the user can see the letters being printed on the display.
- User-friendly design of speech assistant system: The user can see the letters in the display of the speech assistant system. The user can store and delete these letters to make meaningful words and store them permanently in a predefined list, that needs to be pronounced.
- Application-oriented approach for enlisting favorite words: The user is given a list of frequently used words as a part of a favorite list that one uses daily to announce through the speech assistant. This intuitive approach to the project serves as a credible asset to a user who faces a speech disability disorder.
- Translation of words into any language: The user can save any meaningful word to the list. The default setting is English. Four language settings are added: Hindi, Kannada, Tamil, and Telugu. In the online mode, the word in the selected language will be announced/synthesized by Google speech assistant, which holds a very strong connection to the native accent the user desires the word to be announced for fellow listeners. This brings a new dimension to the project.
- Interrupt-oriented switching of online to offline modes: This step is taken particularly for a user who is in online mode and may not have strong Internet connectivity. In such a case, an interrupt-based dedicated program is developed and uploaded so that the user chooses to switch to offline mode to announce the chosen word to fellow listeners immediately. It can pronounce the word in online mode when it finds strong internet connectivity.

4.3 Basic GUI Features

The speech assistant (speechstant) device has a user-friendly GUI so the user can choose the gender of the voice sample, language, and also, we can choose words from predefined lists, and store

letters/words by operating a joystick. The same chosen or stored letter or word is announced for fellow listeners. The speech assistant device uses machine learning and natural language processing libraries incorporated for the pronunciation of words in any language that Google Translate can enlist. It has offline and online modes for the user's benefit. The GUI features are described in Tab. 4. The block diagram of and hardware setup of the joystick module and GUI features are shown in Fig. 12.

Table 4: Basic GUI features description

Scan	In this mode, any input from the dummy tooth set can be seen on the display.
Scan window	This is the general window, where XXXX is input from the four sensors, and Y is output Letter 'a' is stored. This continues until the end of the alphabet. Only when the word is stored does pressing the center of the joystick save the word.
Assistant mode	Any stored word/number/letter can be accessed from the Basic and Emergency file list, and it can be announced.
Info: Information mode	Basic information regarding the movement of the joystick is shown on a page of the user's preference.
Basic	Any word/letter/number that is stored utilizing data acquisition from a dummy tooth/jaw set.
Emergency	The list of predefined words that the user uses regularly can be accessed very fast.
Offline mode	A machine learning library is used for this model. It is the default mode of the speech assistant device. It is designed in such a way that even if the user is not in a Wi-Fi zone if the user wants any particular word to be announced, this mode accomplishes that.

(Continued)

Table 4: Continued

Online mode	Machine learning and a library enabled with natural language processing are enabled in this mode. In this online mode, a user can announce any word in over 60 languages. This mode needs Wi-Fi connectivity. The latency of the mode depends on internet connectivity. Three stages of this mode are: (i) Upload-(.) Single dot (ii) Download-(..) double dot (iii) Speak-(. . .) triple dot
Language select	The four languages that are set for the benefit of the user are: (i) Tamil (ii) Kannada (iii) Telugu (iv) Hindi. These four languages can be accessed only in the online mode. English is the default language of the speech assistant device.
Machine learning libraries	
ES_SPEAK	Every letter/word/number that is stored in the memory or that the user sees in the scan window of the LCD panel is announced by this library. It is an easily compatible Python-based machine learning library that can be easily used to announce letters/words/numbers/in offline mode. It can only use the default language, English, for the announcement of words. In the case of those that are written in English but whose origin is any other language, it will still pronounce with an anglicized accent.
Google translate	Can announce any word/number/letter in online mode (internet connectivity must be available). It is the most robust and complete algorithm that is known to use natural language processing to announce words. The biggest advantage for users is that they can choose the desired language. In this case, a word with Hindi/Kannada/Tamil/Telugu origin that is written in English can be spoken without compromising on the accent. The machine learning library learns from many examples on a daily scale. The chances of error are almost nil.

The hardware setup is connected to Bluetooth, before going to operate the joystick. The input sensor values are analyzed, and display the output based on oral cavity movements in LCD screen

which can be operated by joystick of speechstant device. A block diagram of the joystick module is shown in Fig. 12a.

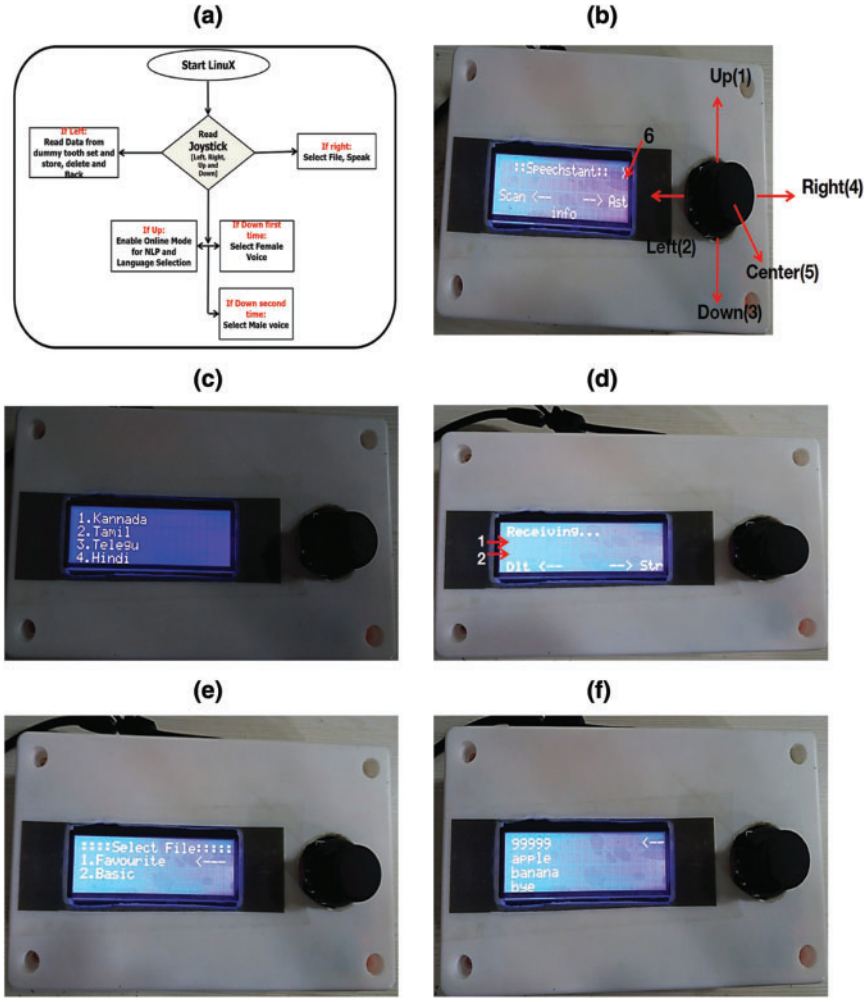

Figure 12: (a) Joystick module (b) Speechstant device (c) Language setting page (d) Scan page (e) Assistant page (f) Assistant page in basic mode

The blue light at point 6 indicates the switch has been read, and it displays the processing status. If point 6 has no sign, it means the user must first connect with the dummy tooth set as shown in Fig. 12b. The joystick up button selects the online mode and the language setting page, in which the up and down buttons of the joystick are used to select the preferred language (Kannada, Tamil, Telugu, Hindi, but the default is English) as shown in Fig. 12c. Similarly, the center and left buttons are used to confirm a selection or to move back and to set the default language is English. The left switch engages the scan mode when the user wants to store a letter to make a sentence communicating with a dummy tooth set over Bluetooth. Fig. 12d shows the scan page. Pointer 1 shows the data are coming continuously, and you can store a letter within 4 s using the right switch. After that, it appears at pointer 2. At that point, you can delete the last letter from the word using the left switch. The down switch selects either male or female voice depending on the disabled person's choice (the default is female). The right switch is

used to select what the user wants to speak from a predefined list. The center switch (point 5) of the joystick is not functional.

When a word is ready, the user stores it to a basic file by clicking the center button followed by the saved message and refreshing it again. To go to the home page, the up switch is used. Fig. 12e shows the assistant page where a user can choose the favorite file (a predefined word that is used frequently or for emergency purposes) or a basic file (the word has already been stored the word by a dummy tooth set) file by up and down button and select it by pressing the center button. Fig. 12f shows the second page of the assistant mode, where a user can choose a word with the up/down button. The arrow is a pointer to the that the user makes decisions about. If the user wants to speak that word, he or she just presses the center button. If the user wants to delete any line, the user must hold the right switch for 2 sec. When the user is speaking, he or she moves to the left switch and holds it to go to the home page. The program must be restarted when the user wants to move from offline mode to online mode.

4.4 Hardware Experimental Results

The importance of the sensor, the variation of the limit switches, the flex, and the potentiometer were displayed on the laptop screen with the help of the Arduino board, as shown in Fig. 13a. The respective output is displayed on LCD through the I2C module, as shown in Fig. 13b, respectively. The produced output sound was heard from the electric loudspeaker according to the sensor data based on the matching with the LUT (see Tab. 4), or from a predefined list stored on the SD card.

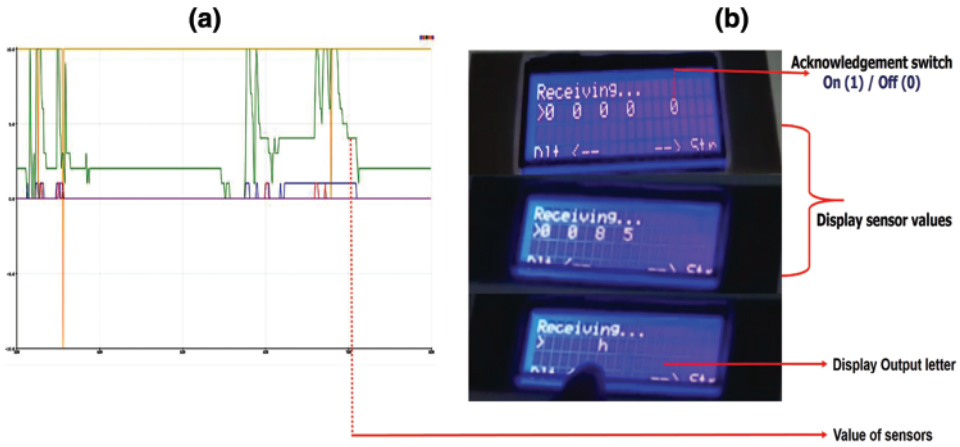

Figure 13: The system for speech production which displays the values of sensor inputs (a) Laptop (b) LCD screen

When the acknowledgment switch was ON {1}, the hardware setup reads the input sensor data. After that, when the acknowledgment switch was OFF {0}, it displayed the corresponding English alphabet based on the match of input sensor data with the LUT, and the same letter is produced loudly by the electric speaker.

The speechstant device displays the sensor values based on the input captured by sensors which are placed in the oral cavity during oral cavity movements. The output screenshots of the LCD screen during the experiment of the proposed hardware setup from the initial page to output letter/words displayed are shown in Fig. 14.

The initial page of the speech assistant device is shown in Fig. 14a. A user moves the joystick down to choose the male voice (the default is the female voice), as shown in Fig. 14b. The sensor input data from the tooth set are matched with the LUT, and the respective letter is displayed in the LCD, once the store button (Str) is pressed. The user can press the delete button (Dlt) to avoid storing and producing the sound. Then the same word is saved in basic mode, and the sound is produced through the speaker in the voice of the chosen gender. The options for a single letter "a" and a three-letter word "add" are shown in Figs. 14c and 14d. The favorite and basic file modes are shown in Figs. 14e and 14f.

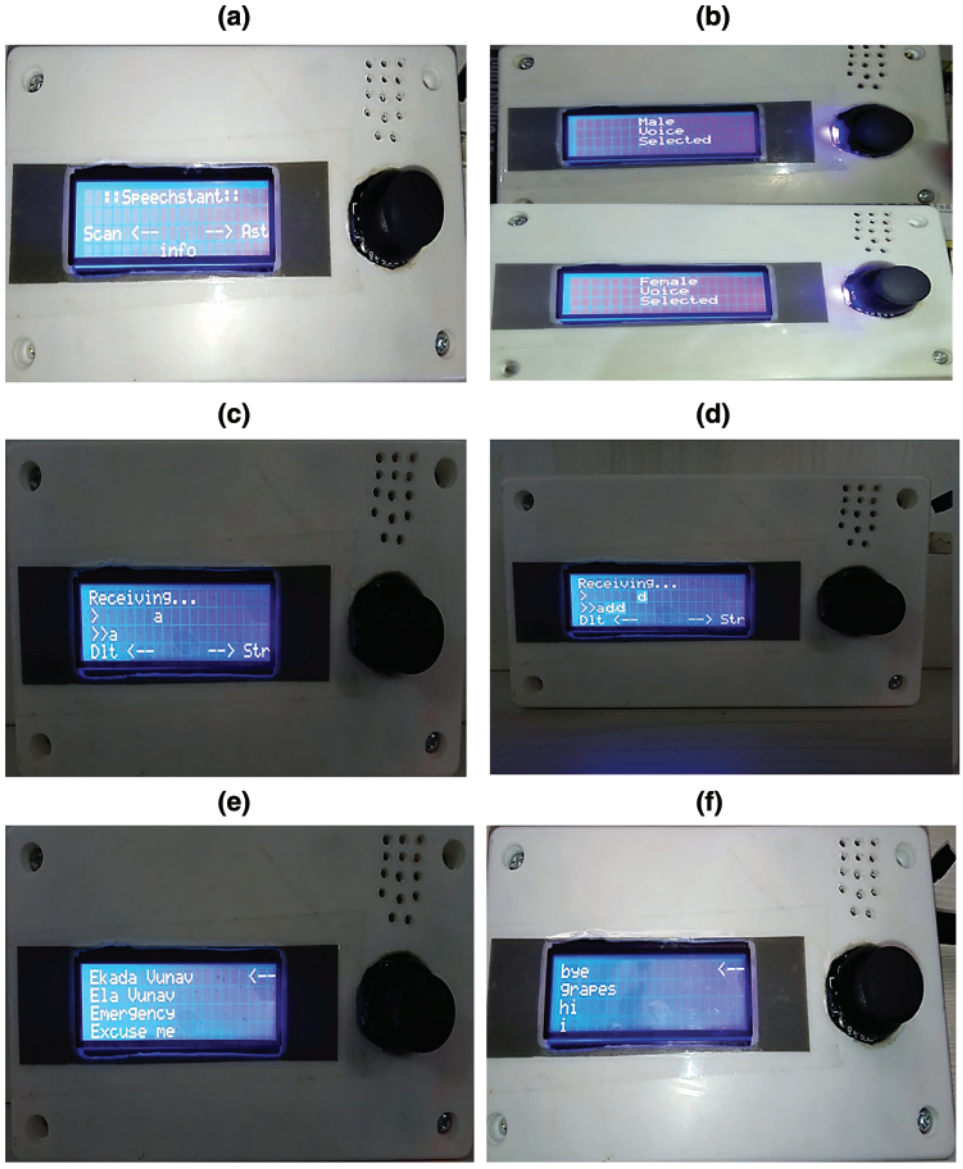

Figure 14: The LCD screen (a) Initial page (b) Gender selection and (c) One letter output (d) Three-letter word output (e) Frequently used word list (f) Predefined list

The output sound waveforms produced based on the interfacing sensor input values with the proposed hardware system are shown in Fig. 15 for both female and male voices. In general, female

voices have a higher pitch (amplitude) when compare to male voices, the same differences can also be observed in Fig. 15. The comparison between natural speech and its synthesized speech waveforms is shown in Fig. 16 and observed that they were almost similar.

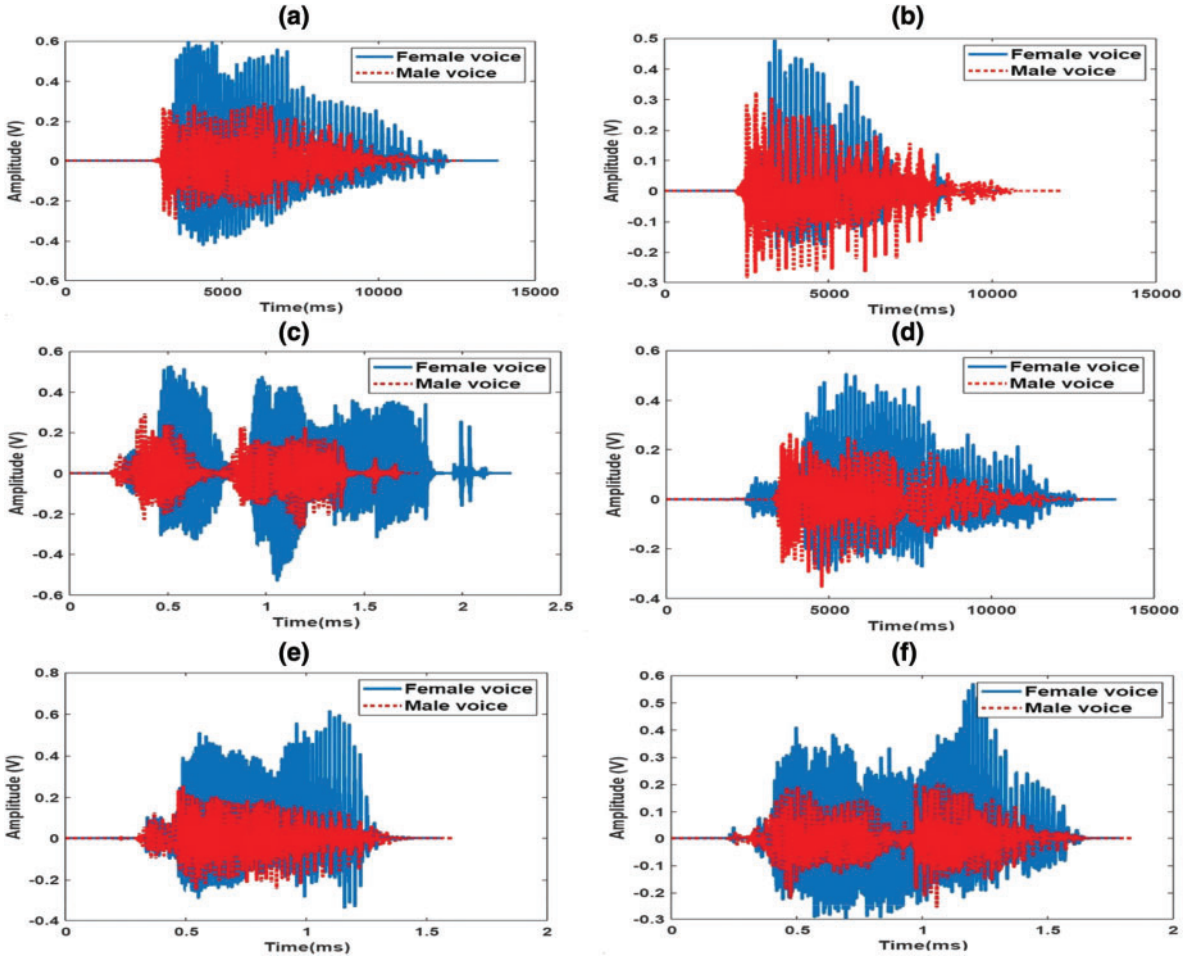

Figure 15: The output waveforms of female and male voices (a) letter "a" (b) "add" [English] (c) "baguna" [Telugu] (d) "banni" [Kannada] (e) "ghar" [Hindi] (f) "panam" [Tamil]

The output sounds produced from the hardware setup based on the sensor inputs have differences in production time. The production time differs based on the average length of the word (number of the letters) is shown in Tab. 5.

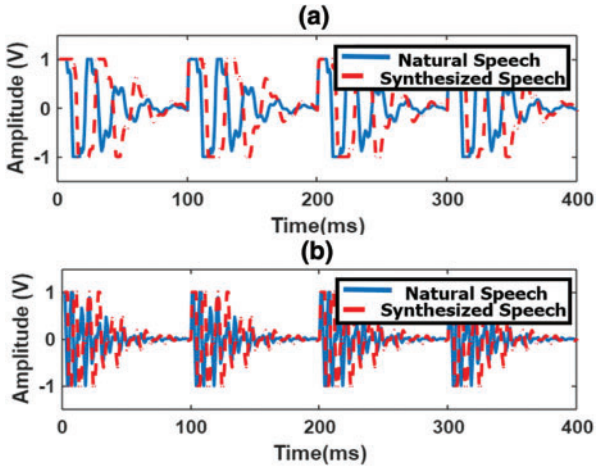

Figure 16: The comparison between natural and its synthesized speech waveforms for (a) /a/ (b) /w/

Table 5: The output sound production time in seconds for letter "a", "add" [English], "baguna" [Telugu], "banni" [Kannada], "ghar" [Hindi], "panam" [Tamil]

Words	Production time (sec)
A	1
Add	3
Baguna	6
Banni	5
Ghar	4
Panam	5

The output sound produced from the input sensor based on oral cavity gestures is validated by a Mean Opinion Score (MOS) which is discussed in the following subsection.

4.5 Mean Opinion Score (MOS)

The opinion score [29] was validated by 20 listeners (10 males and 10 females), all native speakers of British English aged 17–42 years old, who were recruited to participate in the experiment. The listeners had no known speaking or hearing impairments. The test was devised to evaluate the quality of the synthesized speech produced through voice conversion. The opinion score measures and judge the correctly produced sound by the listeners. A five-point scale was used, with five as the best score. The scores from the evaluation test are shown in Tab. 6. Tab. 6a shows the means for correctly identifying the gender of the voice sample, and Tab. 6b shows the means for correctly identifying the stated words over the proposed hardware system.

The overall performance in correctly identifying the speech for all ages of listeners was quite good. Approximately 98% accurately identified the gender, and there was 95% accuracy in identifying the words in the voice samples. The pair of voice samples of the words add, had and pan, span was similar, and they created some issues of perception or quality that left some listeners confused. This contributed to the 95% accuracy in identifying the voice samples of all the words.

Table 6: Mean opinion scores

	(a) Gender		(b) Words of the voice sample						
	Female	Male	add	had	bad	dad	pan	span	team
MOS	4.8	4.9	4.6	4.7	4.8	4.7	4.6	4.8	5

5 Conclusion and Future Work

This paper presents an approach for speech production using oral cavity gestures especially movements of the tongue, lips, and teeth. our motivation is to make communication easier for the speech disabled. Speech disability can occur because of cancer of the larynx, spinal cord injury, brain injury, neuromuscular diseases, or accidents. The four positions of the sensors in the proposed system were based on appropriate articulatory (oral cavity) gestures estimated from the mechanism of human speech, using VocalTractLab and the vocal tract acoustics demonstrator (VT Demo). From the study and analyses of existing vocal tract speech production physiology, we observed the tongue plays a crucial role in speech production. An initial experiment was carried out by listing the positions of oral cavity gestures for respective sound production. It was tested by emerging a GUI using Matlab.

With the reference of knowledge from the initial experiment, the hardware system was verified using an experimental dummy tooth set setup with four sensors, and it produces the speech. The tongue and jaw movements placed in the dummy tooth set were captured by two limit switches, a potentiometer, and a flex sensor. These sensor data from oral cavity movements translate into a set of user-defined commands by developing efficient algorithms that can analyze what is intended and create a voice for those who cannot speak. The output sounds can be heard from an electric speaker, and they are displayed on the screen of the speech assistant device. The system was extended to apply to other languages, such as Hindi, Kannada, Tamil, and Telugu, using a language translator. Based on the choice of speech-disabled, users can select gender voice samples as male or female voice. Those results were validated by a perceptual test, in which ∼98% accurately identified the gender of the voice, and there was ∼95% accuracy in identifying the words in the voice samples. Thus, this system was useful for speech-disabled because of accidents, neuron disorder, spinal cord injury, or larynx disorder to have communication easily.

In future research, time delays can be reduced during sound production. The present work could be extended to generate sequences of words and long/whole sentences. Using the basic facts demonstrated in this study, it might be possible to build a chip design system that wirelessly tracks the movements of the tongue and transmits the sensor data through Bluetooth to a personal computer in which data are displayed and saved for data analysis. Another objective will be to include emotion in the output speech of the proposed system, to communicate to express their thoughts with emotions.

Acknowledgement: The authors thank all the participants who enabled us to validate these output sounds/words.

Funding Statement: The authors would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India for financial support through the scholarship for Palli Padmini, during research work through Visvesvaraya Ph.D. Scheme for Electronics and IT.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. H. Eide and M. E. Loeb, "Data and statistics on disability in developing countries," *Disability Knowledge and Research Programme Executive Summary*, 2005. [Online]. Available: http://www.disabilitykar.net/research/thematic_stats.html.

[2] K. Bunning, J. K. Gona, V. Odera-Mung'ala, C. R. Newton, J. A. Geere *et al.,* "Survey of rehabilitation support for children 0–15 years in a rural part of Kenya," *Disability and Rehabilitation*, vol. 36, no. 12, pp. 1033–1041, 2014.

[3] L. I. Black, A. Vahratian and H. J. Hoffman, "Communication disorders and use of intervention services among children aged 3–17 years: United States, 2012," *NCHS Data Brief*, vol. 205, no. 1, pp. 1–8, 2015.

[4] J. Law, J. Boyle, F. Harris, A. Harkness and C. Nye, "Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature," *International Journal of Language and Communication Disorders*, vol. 35, no. 2, pp. 165–188, 2000.

[5] G. R. Divya, H. Jayamohanan, N. V. Smitha, R. Anoop, A. Nambiar *et al.,* "Primary neuroendocrine carcinoma of the larynx: A case report," *Indian Journal of Otolaryngology and Head & Neck Surgery*, vol. 72, no. 3, pp. 1–4, 2020.

[6] A. J. Venker-van Haagen, "Diseases of the larynx," *Veterinary Clinics of North America: Small Animal Practice*, vol. 22, no. 5, pp. 155–172, 1992.

[7] C. E. Steuer, M. El-Deiry, J. R. Parks, K. A. Higgins and N. F. Saba, "An update on larynx cancer," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 31–50, 2017.

[8] R. T. Sataloff, M. J. Hawkshaw and R. Gupta, "Laryngopharyngeal reflux and voice disorders: An overview on disease mechanisms, treatments, and research advances," *Discovery Medicine*, vol. 10, no. 52, pp. 213–224, 2010.

[9] T. Schölderle, A. Staiger, R. Lampe, K. Strecker and W. Ziegler, "Dysarthria in adults with cerebral palsy: Clinical presentation and impacts on communication," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 2, pp. 216–229, 2016.

[10] L. R. Rabiner and B. H. Juang, in *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1993. [Online]. Available: https://go-pdf.online/fundamentals-of-speech-recognition-l-rabiner.pdf.

[11] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 325–332, 2004.

[12] P. Vijayalakshmi and M. Aarthi, "Sign language to speech conversion," in *Proc. Int. Conf. on Recent Trends in Information Technology (ICRTIT)*, Chennai, India, pp. 1–6, 2016.

[13] A. Katsamanis, E. Bresch, V. Ramanarayanan and S. Narayanan, "Validating rt-MRI based articulatory representations via articulatory recognition," in *Proc. Twelfth Annual Conf. of the Int. Speech Communication Association*, Florence, Italy, pp. 2841–2844, 2011.

[14] E. C. Lu, T. H. Falk, G. Teachman and T. Chau, "Assessing the viability of a vocal cord vibration switch for four children with multiple disabilities," *The Open Rehabilitation Journal*, vol. 3, no. 1, pp. 55–61, 2010.

[15] B. Denby, T. Schultz, K. Honda, T. Hueber and J. M. Gilbert, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[16] M. Scudellari, "Brain implant can Say what you're thinking," *IEEE Spectrum*, April, 2019. [Online]. Available: https://spectrum.ieee.org/the-human-os/biomedical/devices/implant-translates-brain-activity-into-spoken-sentences.

[17] D. Arsh Shah, "16-year-old invents a breath enabled talking device to help the speech impaired," thebetterindia.com, 2014. [Online]. Available: https://www.thebetterindia.com/12730/now-speech-impaired-people-talk-simple-device-arsh-dilbagi-aac/.

[18] K. U. Menon, R. Jayaram and P. Divya, "Wearable wireless tongue controlled assistive device using optical sensors," in *Proc. 2013 Tenth Int. Conf. on Wireless and Optical Communications Networks (WOCN)*, Bhopal, India, pp. 1–5, 2013.
[19] K. U. Menon, R. Jayaram, D. Pullarkatt and M. V. Ramesh, "Wearable wireless tongue-controlled devices," *United States Patent*, Washington, DC, US 9,996,168. 2018.
[20] P. Padmini, D. Gupta, M. Zakariah, Y. A. Alotaibi and K. Bhowmick, "A simple speech production system based on formant estimation of a tongue articulatory system using human tongue orientation," *IEEE Access*, vol. 9, pp. 4688–4710, 2020.
[21] P. Padmini, S. Tripathi and K. Bhowmick, "Sensor based speech production system without use of glottis," in *Proc. Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, pp. 2073–2079, 2017.
[22] P. Birkholz, "Vocaltractlab 2.1 user manual," *Technische Universität Dresden*, 2013. [Online]. Available: https://www.vocaltractlab.de/.
[23] M. Huckvale, "VTDemo-vocal tract acoustics demonstrator," *Computer Program University College London*, 2009. [Online]. Available: https://www.phon.ucl.ac.uk/resource/vtdemo.
[24] G. N. Clements, "The geometry of phonological features," *Phonology*, vol. 2, no. 1, pp. 225–252, 1985.
[25] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.
[26] Switch, "robu. in," *Macfos Private Limited*, 2021. [Online]. Available: https://robu.in/product/tact-switch-kw11-3z-5a-250v-micro-switch-round-handle-3-pin-n-o-n-c-for-3d-printers/?gclid=Cj0KCQjwwY-LBhD6ARIsACvT72NPrWnyvU4qUbS9oFYcW_k5Xzti7w2aLGAeAtrk96iC-FQHA88G6FwaAnpVEALw_wcB.
[27] Flex sensor 2.2", "Generation robots," *Spectra Symbol Technology*, 2021. [Online]. Available: https://www.generationrobots.com/en/401948-flex-sensor-22.html.
[28] Potentiometer, "Indiamart," *Pankaj Potentiometers Private Limited*, 2021. [Online]. Available: https://www.indiamart.com/proddetail/single-turn-wire-wound-potentiometer-4714719712.html.
[29] R. C. Streijl, S. Winkler and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 231–227, 2016.