IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 12, 2020, accepted December 26, 2020, date of publication December 29, 2020,
date of current version January 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048076

A Simple Speech Production System Based on Formant Estimation of a Tongue Articulatory System Using Human Tongue Orientation

PALLI PADMINI[1], DEEPA GUPTA[2], MOHAMMED ZAKARIAH[3], YOUSEF AJAMI ALOTAIBI[4], (Senior Member, IEEE), AND KAUSTAV BHOWMICK[5], (Member, IEEE)

[1] Department of Electronics & Communication Engineering, Amrita School of Engineering, Bengaluru 560035, India
[2] Department of Computer & Science Engineering, Amrita School of Engineering, Bengaluru 560035, India
[3] Research Center, College of Computer and Information Science, King Saud University, Riyadh 11451, Saudi Arabia
[4] Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[5] Department of Electronics and Communication Engineering, PES University, Bengaluru 560085, India

Corresponding author: Deepa Gupta (g_deepa@blr.amrita.edu)

This work was supported in part by the Research Groups Program (Research Group Number RG-1439-033), under the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, and in part by the Government of India's Visvesvaraya Ph.D. scheme through the scholarship for Palli Padmini.

ABSTRACT An algorithm for a potentially non-obtrusive speech production system was developed and characterized. The algorithm is primarily based on the articulation of the human tongue referred as tongue articulatory system (TAS) and was cascaded with a previously developed laryngeal model. We developed and optimized statistical formulae for formants of vowels and consonants and studied the model for different ages and genders. The difference between the formant frequencies obtained using both the established vocal tract system and proposed cascaded system was found to be $< 5\%$. The proposed model shows the significance of the articulatory nature of the tongue in human speech production. An algorithmic speech synthesizer was developed, and its output was matched with original speech signals for English vowels and consonants with an Normalized Root-Mean-Square deviation error (NRMSE) of < 0.15ms. Further, an experimental implementation of the developed algorithm was done, with flex-sensors emulating the tongue in an artificial oral cavity. The experimental test results further confirmed the effectiveness of the algorithm, revealing interesting features under tolerance analyses. This idea relates to a means for compensating for a whole or partial loss of speech. Such a model can be useful to interpret speech for tracheostomised patients who have undergone larynx surgery, speech-disabled due to accidents or voice disorders, medical rehabilitation and for robotics.

INDEX TERMS English vowels and consonants, formants, Laryngeal system, oral cavity system, sensors, tongue, vocal tract system.

I. INTRODUCTION

Speech is the most common method to communicate human thoughts. Unfortunately, an estimated ~5–11% of people have speech disorders and cannot rely on natural speech for communication [1]–[4]. People with speech disorders have problems creating or forming the vocal sounds needed to communicate with others. Common speech disorders include articulation disorders, phonological disorders, disfluency, and voice disorders, as a result of damage to parts of the brain or nerves, cleft palate or other problems with the palate, overuse of the vocal cords, ulcers on vocal cords, and laryngeal webs [5]. A lack of verbal communication affects social participation, education, and employment because of limited direct interaction. Therefore, the development of a mechanism to facilitate straightforward communication through synthesized speech is necessary to benefit patients with speech disorders. Recently, the synthesis of speech based on articulatory gestures has caught the attention of researchers, who mostly use a clinical approach. Understanding the relationship between articulatory features and acoustic signals [2], [3] is essential to produce acoustic speech by solving the articulatory inversion problem. A wide range of research has been conducted globally on many speech synthesis systems for people with disabilities. For instance, an electrolarynx [6], which is placed on a neck strap, enables

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

the user to modulate muscles to create speech by moving their articulators, including the lips, jaw, tongue, and teeth. Silent speech systems [7] were made to record the activity of the articulatory system, including the facial muscles, using electromagnetic articulography (EMA) signals to detect the synthesis of speech signals that are spoken silently. An articulatory speech synthesizer as a generic acoustic model [8], was used to determine the acoustic performance of articulatory mapping by describing the articulatory trajectories of the jaw, lips, tongue body, and tongue tip. Vocal cord vibration switches [9] were positioned on the throat and capture sensor signals, which are sent to an iPod through a Bluetooth transmitter to detect the periodic vibrations associated with vocalizations. Also, in silent sound technology, an electromagnetic sensor is attached to the face and records pulses, which were converted to speech [10]. The TALK device was made to convert breath to speech [11], and the tongue and ear interface [12] is a wearable system that captures the tongue and jaw movements that are used for speech recognition.

Few of these systems follow the vocal tract model, involving possible implants at the glottis, and they are equally dependent on the larynx and oral cavity. The remaining systems [7], [8], [10]–[12] synthesize speech based on image processing or articulatory sensor data, making them more complicated to use. An array of magnetic sensors was used to wirelessly track the movements of the tongue by detecting the position of a permanent magnetic tracer secured to the tongue. Tongue movements have been translated into different commands and used to access a computer or control a motorized wheelchair, phone, TV, or robotic arm [13], [14]. The tongue, which has been clinically characterized thus far, for vowel production [15], [16], has been considered as the main player in speech production by the oral cavity, in the present work. Herein, we have conceptualized a tongue-based human speech producing system for developing a speech production solution for people who have lost their voice due to accidents, larynx disease [17]–[19], larynx disorders [20], dysarthria, or cerebral palsy [21] and for medical rehabilitation and for robotics.

Thus far, only some experimental medical studies have been reported on understanding tongue contours during speaking, e.g., ultrasound imaging measurements of the x and y coordinates of the tongue and its curvature positions have been reported [22] to represent the tongue during speech production. Other work has estimated the formant frequencies of vowels in an articulatory model based on the combination of the jaw, tongue, and larynx [23]. Vowel formant frequency values were experimentally estimated using the recorded speech of 18 healthy adults, which were correlated with their tongue curvatures, which were obtained using ultrasonography to analyze the resonance mechanisms of the oral vocal tract system [24]. Thirteen healthy female speakers were studied to qualitatively estimate a relationship between the tongue's x-y coordinates and formant frequencies [25], and it was concluded that the first formant frequency depends upon the height of the tongue, and the second formant depends on the advancement of the tongue. Formant frequencies along with the tongue and jaw positions of two female singers were studied using X-ray during the articulation of /a/, /i/, and /u/ to determine the jaw articulatory parameter that relates to pitch [26]. Acoustic and electromyography (EMG) analyses were performed during 12 Dutch vowel articulations repeated 30 times, confirming the lesser significance of lips in producing the first formant [27]. The articulation of English vowels by five speakers, related to a two-dimensional formant space comprised of individual tongue points, were observed to correlate and predict points to define the accuracy of the model [28]. The characteristics of 3D vocal tract geometry for approximant consonant sounds were studied to analyse the articulatory-acoustic models, using magnetic resonance imaging (MRI) and electropalatography (EPG) data [29]. Also, the articulation of nasal consonants was studied using a database of 1200 words recorded by six speakers, including three male and three female members [30]. The recording of consonant syllables, i.e., plosives (/b/, /p/, /d/, /g/, /k/, /t/) of English phoneme by a male speaker in a carrier phrase using the spectrogram perceptual process may entail continuous tracking of vocal tract resonances [31]. Although the maneuvers of the tongue in speech production have been mapped, a ubiquitous statistical model for tongue-based vowel and consonant production and speech reproduction has not been formulated yet.

Studies in the literature, based on qualitative data, have proposed that the first formant (F_1) is inversely proportional to the height of the tongue body, and the second formant (F_2) frequency is related to the size of the frontal oral cavity or the degree of tongue advancement based on X-ray photographs showing the positioning of the tongue and lips [24]–[28]. We studied the experiments conducted using ultrasonography, EMA, and X-ray, as discussed in the above literature, and concluded that tongue positions majorly correlate with the first two formants of the vowels for speakers. The formant frequencies based on the experiments in the studied literature were speaker-dependent and varied by gender and age. In our work, starting with the accumulated results on vowels, we have proposed optimized statistical formulae for vowel formant frequencies and extended the work to consonants, with all the research based on tongue movement-mapping during vowel and consonant pronunciation. The proposed statistical model for the tongue-based oral cavity has been cascaded with a laryngeal model, and we conducted a detailed comparison with speech produced by the vocal tract model. The suitability of the proposed algorithm, based on the formant expressions, was used to generate vowel and consonant sounds for various age groups and both genders: males, females, and children aged nine years old. Thus, focusing the work primarily on tongue gestures, via the proposed model in this article, simplifies making a speech production device.

Overall, the present work shows the statistical foundation of a human-tongue-based speech production system, which may be easily used in a wearable system, potentially

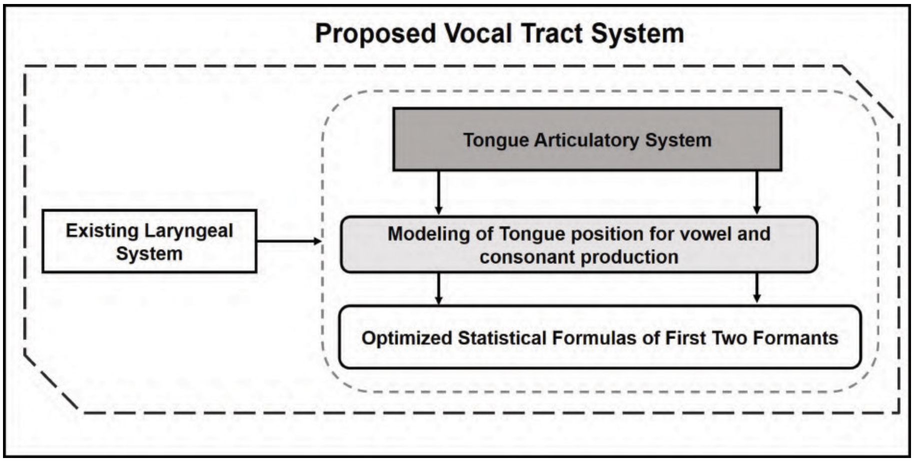

Proposed Vocal Tract System
Tongue Articulatory System
Existing Laryngeal System
Modeling of Tongue position for vowel and consonant production
Optimized Statistical Formulas of First Two Formants

FIGURE 1. Schematic representation of the proposed vocal tract system with a simplified tongue articulatory system.

without physical intrusions. The major contribution presented in subsequent sections is outlined below.

- The proposition of an optimized statistical relation for the first two formants of English vowels and consonants using human tongue gestures to define an age and gender independent speech production system.
- Characterization of a complete model with the proposed tongue-based articulatory system and a known laryngeal model cascaded together.
- Validation of results of the proposed system using the existing vocal tract system and acoustic error (E) based on the formants from frequency response and Normalized Root-Mean-Square Error (NRMSE) from synthesized sound using a formant synthesizer.
- Experimental confirmation of the developed theory with the dummy tongue model setup using flex sensors arranged as a human tongue to produce tongue-like movements for sound production whose outputs were sounded and displayed through an Arduino based assembly, mobile screen through the serial monitor app and verified with a speech from the electric speaker.

Section II describes the flow of the proposed system and formulation of the formant frequencies of the oral cavity system. Then the results and discussion are described in Section III, and Section IV describes the experimental hardware setup, followed by a conclusion and future research expectations in Section V.

II. METHODS

The proposed methodology aims to produce a speech sound using the oral cavity system formants, which are estimated statistically based on the orientation of the tongue. After establishing the relevant expressions, the proposed tongue-based system was cascaded with a known laryngeal system [32], as shown in Fig. 1.

A. PROPOSED TONGUE ARTICULATORY SYSTEM

Here, in Section II.A, the estimation and optimization of the tongue-position-based English vowel and consonant pronunciation are presented. The proposed oral cavity system is particularly based on tongue orientation characteristics during the articulation of the English alphabet. Thus, the proposed oral cavity system is called the ‘‘tongue articulatory system (TAS)’’.

1) MODELING OF TONGUE POSITION FOR VOWEL PRODUCTION

From the literature reviewed in Section I, the first two formants of oral cavity system formants are estimated using the concept from the literature, which is inversely proportional to tongue height and tongue advancement, respectively, during vowel articulation. We conducted statistical estimations by mapping tongue orientation characteristics by adopting the vocal tract synthesizer [33], [34], and vowel space theory [35], [36]. In Fig. 2, vocal tract shapes and quadrilaterals are shown in pairs, representing each vowel.

Fig. 2a shows the articulatory targets that correspond to English vowels adopted from a previously reported model, VocalTractLab [33]. The same pattern, which is in quadrilateral shape, is replicated in vowel space theory, in which the horizontal axis indicates tongue advancement (l) (e.g., front, central, or back) which describes the tongue being raised and slant line as tongue height (h) (e.g., close, mid, and open) during vowel articulation. The vowel space was measured in centimeters (cm) following the standard vowel quadrilateral, as shown in Fig. 2b, with the coordinate point labeled (l,h) to represent the tongue shape and position required for the articulation of each vowel sound. Using l and h values, as given in Fig. 2b, we formulate the formulae for the tongue articulatory system formants, which are discussed in the following subsection.

a: OPTIMIZED STATISTICAL FORMULAE OF THE FIRST TWO FORMANTS OF THE TONGUE ARTICULATORY SYSTEM FOR THE VOWELS

The first formant, denoted as F_1^o, has a value that is inversely proportional to the tongue height (h) for the production of the

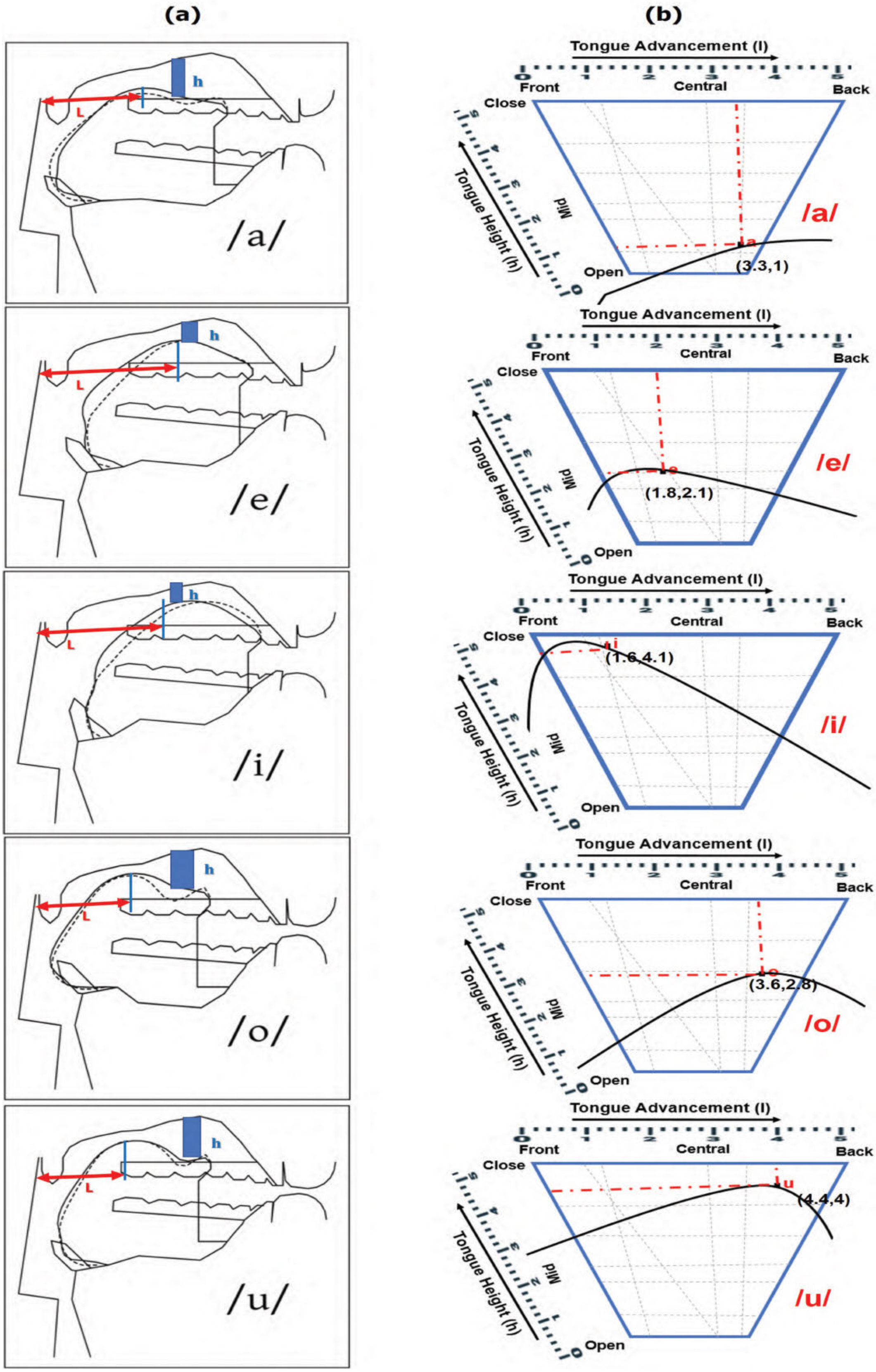


FIGURE 2. **(a) Vocal tract shapes and (b) Relationship between the articulatory setting in terms of tongue height and advancement for English vowels.**

vowel sound, as given in Eq. (1),

$$F_1^o \propto \frac{1}{h} \tag{1}$$

The second formant, denoted as F_2^o, has a value that is inversely proportional to the tongue advancement (l) during the production of the vowel, as given in Eq. (2),

$$F_2^o \propto \frac{1}{l} \tag{2}$$

The oral cavity was considered a tube model and was assumed to be a resonator. Post-study of the first two formants of the oral cavity system based on the tongue positions for vowels from Fig. 2b, a proportionality constant c and scalar constants β_1 and β_2 were introduced into Eqs (1) and (2) obtaining,

$$\hat{F_1^o} = \beta_1 \frac{c}{h} \tag{3}$$

$$\hat{F_2^o} = \beta_2 \frac{c}{l} \tag{4}$$

where β_1 and $\beta_2 \in R$, and c is considered the speed of sound (34,300 cm/sec). The next step was to identify the closest constant values of β_1 and β_2 with the responses of the tongue articulatory system formants provided in Eqs (3) and (4), which are approximately similar to the existing oral cavity system formants based on experimental values using MRI and X-ray [24], [34].

To estimate the values of β_1 and β_2, loss function was calculated between the formants of the estimated system [24], [38] and tongue articulatory system from Eq. (3) and (4) using the Mean Squared Error function to calculate the loss, which is given in Eq. (5).

$$J(\beta_1, \beta_2) = \min_{\beta_1,\beta_2} \frac{1}{2} \sum_k [(F_{1k}^o - \hat{F_{1k}^o})^2 + (F_{2k}^o - \hat{F_{2k}^o})^2] \tag{5}$$

where k defines English vowels /a/, /e/, /i/, /o/, /u/; F_1^o, F_2^o and $\hat{F_1^o}$ and $\hat{F_2^o}$ are the two formants of estimated and tongue articulatory system formants.

We applied gradient descent method [37] to find β_1 and β_2 by using the following steps:

1) Initially we let $\beta_1 = 0$ and $\beta_1 = 0$ and η be our learning rate, This controls how much the value of changes β_1 and β_2 with each step. η could be a small value like 0.0001 for good accuracy.
2) The partial derivative of the loss function $J(\beta_1, \beta_2)$ was calculated for β_1 and β_2 and updated the current value of β_1 and β_2 is using an Equation (6).

$$\beta_j := \beta_j - \eta[\frac{\partial}{\partial \beta_j} J(\beta_1, \beta_2)] \tag{6}$$

where j = 1 and 2.

3) We repeated this process until $\|\nabla J(\beta_1, \beta_2)\| < \epsilon \sim 0.001$.

The optimal values for $\beta_1 \sim 0.02$, $\beta_2 \sim 0.1$ pair, was obtained by the minimizing function shown in equation 5, graphically depicted as a red trail in figure 3. The optimized estimated formulae for $\hat{F_1^o}$ and $\hat{F_2^o}$ of the vowels are expressed in Eqs (7) and (8), respectively, after substituting the value of β_1 and β_2 from Eqs (3) and (4) using gradient descent.

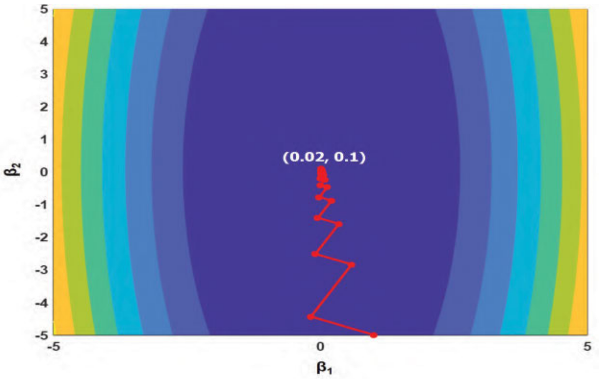


FIGURE 3. Using gradient descent to find the optimum values of β_1 and β_2 of the tongue articulatory system formants for vowels.

TABLE 1. Tongue articulatory system formant values for vowels.

S.No	Vowels	Existing vocal tract system values [5]		Estimated formant values of tongue articulatory system without larynx	
		First formant F_1	Second formant F_2	First formant $\hat{F_1^o}$	Second formant $\hat{F_2^o}$
1	/a/	896.3	1308.1	762.2	1008.8
2	/e/	415.2	1978.5	343	1905.5
3	/i/	222.8	2317	142.9	2413.7
4	/o/	360	1091	228.6	902.6
5	/u/	255	1100	155.6	857.5

$$\hat{F_1^o} = (0.02)\frac{c}{h} \tag{7}$$

$$\hat{F_2^o} = (0.1)\frac{c}{l} \tag{8}$$

Table 1 lists the average values of 60 speakers, which are considered as formant frequencies for the vowel of the existing vocal tract system [5] and the tongue articulatory system formant values after substituting the values of tongue height (h) and advancement (l) from Fig. 2b into Eqs (7) and (8).

Table 1 shows the formant values of the existing vocal tract system includes the larynx and oral cavity with the tongue articulatory system. The differences between the formant values are due to the absence of a laryngeal section in the tongue articulatory system, and they signify the larynx in the existing vocal tract system.

2) MODELING OF TONGUE POSITION FOR CONSONANT PRODUCTION

The relationship between tongue height and advancement for consonants has not been well-reported in the literature, however, the study of articulatory gestures of consonant production can be used to understand the tongue positions and movements [39], [40]. Based on the aforesaid information, we established a relationship between the tongue height (h)

TABLE 2. Statistical formulae for the first two formants of the tongue articulatory system for consonant groups.

Consonant groups	Tongue articulatory system formants formulae		Terminology
	First formant $\hat{F_1^o}$	Second formant $\hat{F_2^o}$	
Approximants (/j/, /w/)	$\hat{F_{1a}^o} = \gamma_1 \frac{c}{h}$	$\hat{F_{2a}^o} = \gamma_2 \frac{c}{l}$	c = speed of the sound = 34300cm/s h =height of the tongue in cm l = tongue advancement in cm Duration = duration of voiced and unvoiced in msec i.e., 0-20ms for /p/, /t/, /k/; 60-100ms for /b/,/d/,/g/ and 40ms for nasals, respectively. B= burst release in milliseconds, defines the pressure that has been rising behind the obstruction
Plosives (/p/, /b/,/t/, /d/, /k/, /g/)	$\hat{F_{1p}^o} = \gamma_1 \frac{c*B}{Duration*h}$	$\hat{F_{2p}^o} = \gamma_2 \frac{c}{l}$	
Nasals (/m/, /n/)	$\hat{F_{1n}^o} = \gamma_1 \frac{c*B}{Duration*h}$	$\hat{F_{2n}^o} = \gamma_2 \frac{c}{l}$	

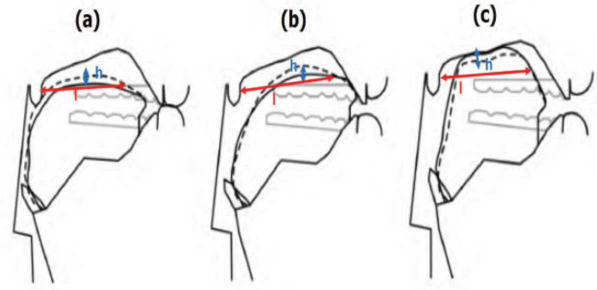


FIGURE 4. Different tongue gestures for the voiced plosives consonants (a) /b/ (Labial), (b) /d/ (Alveolar), and (c) /g/ (Velar) [33].

and tongue advancement (l) in quadrilateral shape for consonants in a way similar to that of the vowels as discussed in Section II.A.1. The statistical formulae of oral cavity formants for consonants were obtained and optimized by using a gradient descent method.

The consonants are described and differentiated using voice-place and manner system [41], [42], based on which, they are divided into five different groups: approximants, plosives, fricatives, affricatives, and nasals. In this section, the approximants, plosives, and nasals groups have been studied, and the remaining two groups follow the same as plosives sounds, only differing by a degree of constriction at the place of articulation [41], [42].

Approximants (/j/ and /w/) are phonetically vowels but phonologically consonants [29]: phonetically they are pronounced as /i:/ and /u:/ but a little bit shorter. English language has six plosive consonants, /p/, /b/,/t/, /d/, /k/, /g/ out of which /p/, /t/, and /k/ are voiceless and /b/, /d/, and /g/ are normally voiced [33]. The position of the tongue during articulation of voiced plosive consonants is shown in Fig. 4.

The nasal consonants are /m/ and /n/, in which there's a closure in the oral cavity and passing of air through the nasal cavity. For this nasal sound production, both oral and nasal cavities are equally considered. The articulation of /m/ and /n/ is brought about by blocking the oral passage, lowering the soft palate, and tongue tip touching the closed teeth line, respectively. The nasals can also be sounded, through the oral cavity even by keeping the nasal cavity closed, but with some loss of clarity and quality of sound. These actions get summarized with the major contribution of tongue height and advancement as this study focuses on oral cavity gestures, especially the tongue. The relationship between the articulatory setting of the tongue in terms of height and advancement for English consonants is shown in Fig.5.

Thus, based on studying tongue positions during the articulation of each consonant sound, the relationship was evaluated in quadrilateral shape with tongue height (h) and tongue advancement (l) on the vertical and horizontal axes, respectively, for the English consonant groups of approximants, plosives, and nasals. Using the relationship between l and h values, as discussed in Fig. 5, we formulated the formulae for the tongue articulatory system formants which are discussed in the following subsection.

a: OPTIMIZED STATISTICAL FORMULAE OF THE FIRST TWO FORMANTS OF THE TONGUE ARTICULATORY SYSTEM FOR THE CONSONANTS

The acoustic properties of consonants have to lead to a statement from the studied literature that the first and second formants are affected by the size of the constriction, manner of articulation (tongue height) defined burst (sudden release of air), the position of the tongue, and voiced or unvoiced sound and place of articulation (tongue advancement). The statistical formulae for the first two formants ($\hat{F_1^o}$ and $\hat{F_2^o}$) of the oral cavity system of approximants, plosives, and nasals are given in Table 2.

Similarly, as discussed for vowels in Section II.A.1.a, we used gradient descent to find the optimum point of scalar constants γ_1 and γ_2, which gives the minimum error between the proposed oral cavity formants of consonants, i.e., approximants, plosives, and nasals, separately with the existing vocal tract system formants [5]. Thus, the optimum point of (γ_1, γ_2) are $\sim$ (0.02, 0.1), (1, 0.2), and (0.1, 0.33), respectively, for approximants, plosives, and nasals using gradient descent to obtain minimum error, as shown in Fig. 6.

The scalar constants γ_1, γ_2 were substituted and values of l, h, and B in statistical formulae given in Table 2 to calculate the first two formants $\hat{F_1^o}$ and $\hat{F_2^o}$ of the tongue articulatory system for approximants, plosives, and nasals. Table 3 shows

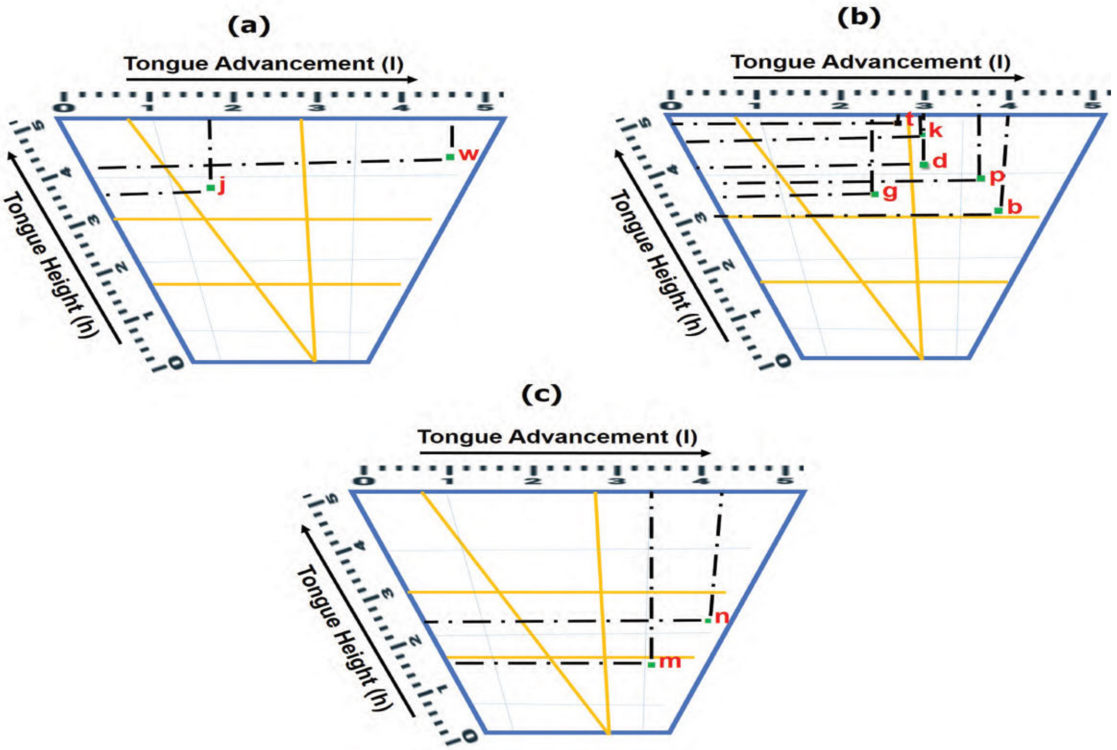


FIGURE 5. Relationship between (*h*) and (*l*) for English consonants (a) approximants /j/ and /w/ (b) plosives /p/, /b/, /t/, /d/, /k/, and /g/, and (c) nasals /m/ and /n/.

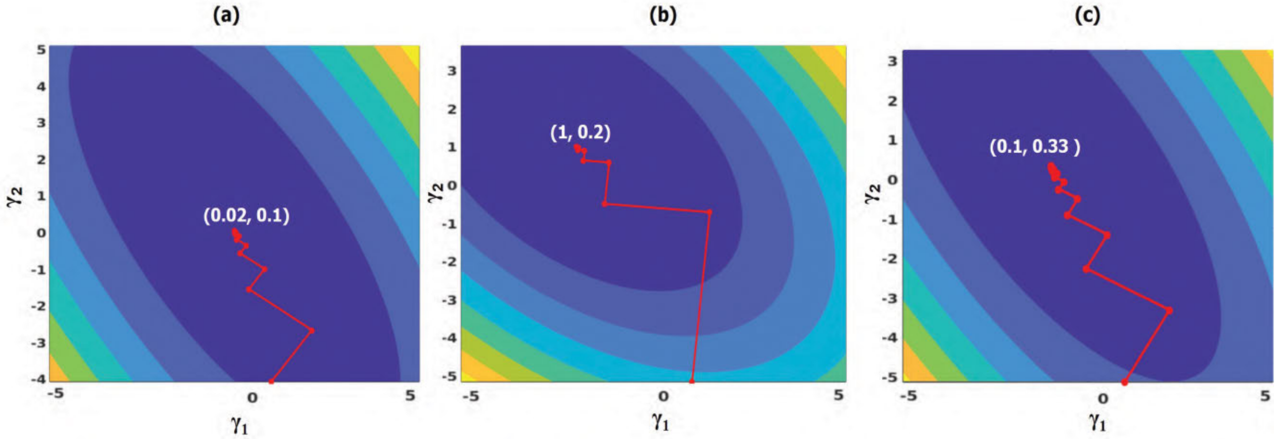


FIGURE 6. Using gradient descent to find the optimum values of γ_1, γ_2 of the proposed oral cavity system formants for (a) approximants, (b) plosives, and (c) nasals.

a comparison of the existing vocal tract system and tongue articulatory system for approximants, plosives, and nasals. The observed difference is due to the absence of a laryngeal section in the proposed system, so far.

Tongue articulatory system formants were derived for approximants, plosives, and nasals, A similar process was followed to derive the formants for remaining consonants groups like fricatives and affricatives, which differ only due to the degree of constriction when compare to plosives. The difference between the formant values signifies that the existing vocal tract system includes the larynx and oral cavity, whereas the tongue articulatory system formants contain only tongue articulation as a physical parameter.

Having established the formants for the complete set of vowels and consonants and by using the aforesaid results, we report a novel method for quantifying speech articulation and suggest that the resonance systems of the first two formants of the tongue articulatory system are distinct and independent of age and gender. Therefore, the tongue is an important articulator and plays a vital role in speech production.

TABLE 3. Formants of the vocal tract and tongue articulatory system for consonants: approximants, plosives, and nasals.

Consonants	Tongue height (h) cm	Tongue advancement (l) cm	Burst release (B) ms	Vocal tract system values [5]		Estimated formant values of tongue articulatory system without larynx	
				F_1	F_2	$\hat{F_1^o}$	$\hat{F_2^o}$
Approximants							
/j/	3.5	1.7	–	300	2450	196	2017.64
/w/	4	4.8	–	250	750	196	857.5
Plosives							
/p/	3.4	3.6	1	252	1943	245	1905.5
/b/	3	4	1	204	1811	190.5	1715
/t/	3.9	2.6	2	350	2775	295.6	2638.4
/d/	4.1	3	2	275	2504	137.7	2268.6
/k/	4.8	2.9	1.6	243	2488	285.8	2365.5
/g/	3.7	2.4	1.6	215	2994	121.9	2858.3
Nasals							
/m/	1.5	3.6	5	543	2519	457.3	2078.8
/n/	2.3	4.3	6	411	2561	372.8	2540.7

B. PROPOSED VOCAL TRACT SYSTEM WITH A SIMPLIFIED TONGUE-BASED ORAL CAVITY

A simplified view of the existing vocal tract system (traditional method) [5] for speech production, as seen in Fig. 7a. The existing vocal tract model contains the lungs (glottal source) and larynx (laryngeal), and oral cavity as a single tube. The lungs act as a power supply and provide airflow to the larynx. The larynx modulates airflow from the lungs and provides either a periodic puff-like or a noisy airflow source. Thus, the output gives the modulated airflow by spectrally shaping the source. The formulation of the proposed vocal tract system by cascading the simplified tongue-based oral cavity system (tongue articulatory system) with the laryngeal system is shown in Fig. 7b. Firstly, the system was used to obtain formants for male speakers, as study literature defines the formant value of the laryngeal system only for male speakers [32]. The laryngeal system formants for female and children speakers are not reported in the literature. Thus, to obtain female and child formants for the proposed system, we verified and used an estimated relationship between the male, female, and nine-year-old children formants [43].

The transfer function of the formant frequencies of the existing vocal tract system is given by a second-order all-pole system [44] expressed as $V(z)_k$ in Eq. (9). The first two formants of the existing laryngeal system (denoted as F_1^L and F_2^L) are 110 Hz and 170 Hz, respectively, identified on vocal fold tissues for male speakers using videostroboscopy discussed in [32]. The transfer function of the proposed formant frequencies of the laryngeal system and tongue articulatory system is given by a second-order all-pole system [44] expressed as $L(z)_k$ and $\hat{O}(z)_k$ in Eqs (10) and (11).

$$V(z)_k = \prod_{i=1}^{2} \frac{1}{1 - 2.(A_1).cos(2.\pi.F_{ik}.T).z^{-1} + (A_2).z^{-2}} \quad (9)$$

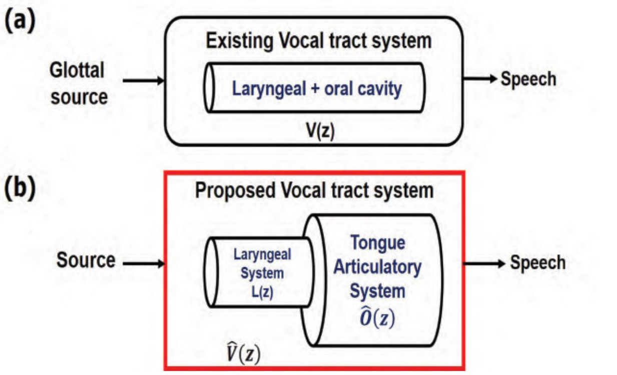


FIGURE 7. Block diagram of (a) an existing vocal tract system and (b) the proposed vocal tract system.

$$L(z)_k = \prod_{i=1}^{2} \frac{1}{1 - 2.(A_1).cos(2.\pi.F_{ik}^L.T).z^{-1} + (A_2).z^{-2}} \quad (10)$$

$$\hat{O}(z)_k = \prod_{i=1}^{2} \frac{1}{1 - 2.(A_1).cos(2.\pi.\hat{F_{ik}^o}.T).z^{-1} + (A_2).z^{-2}} \quad (11)$$

where A_1 and and A_2 are coefficients given in Eq. (9) - Eq. (11), whose values are $exp(-\pi.B_i.T)$ and $exp(-2.\pi.B_i.T)$, respectively, k corresponds to the successive English vowels and consonants. F_i, F_i^L and $\hat{F_i^o}$ denote the formant frequencies of existing vocal tract system, laryngeal and tongue articulatory system, respectively, B_i denotes the bandwidth with values 130 Hz and 70 Hz, and T denotes the fundamental period of approximately 1.2 millisec.

Here, the transfer functions of both the laryngeal system $(L(z)_k)$ and tongue articulatory system $(\hat{O}(z)_k)$ are derived in z-transform (frequency domain). Therefore, the cascaded systems are convolved to obtain the response of the proposed vocal tract system for male speakers $(\hat{V}(z)_k)$, as given in Eq. (12), as shown in Fig. 7b.

$$\hat{V}(z)_k = L(z)_k * \hat{O}(z)_k \quad (12)$$

Thus, by using the formants of the proposed vocal tract system for male speakers obtained from Eq. (12), we derived the formants of a proposed system for female and child speakers. The male, female, and children generally differ significantly in their average vocal tract length for the formant frequencies [5]. For this reason, the same sound is usually represented by different formant frequencies in males, females, and children [43]. The female and child speaker formants are, on average, 17% and 25% are higher than males [5]. The first two formants of female and children formant frequencies are about 12%, and 17%, which are 32% and 37% higher, respectively, than those of male adults [45]-[47]. We experimented by collecting 20 English alphabet sound samples from each male, female, and child speaker to justify the relationship. The formants obtained from the recorded samples were considered actual formants, and the formants

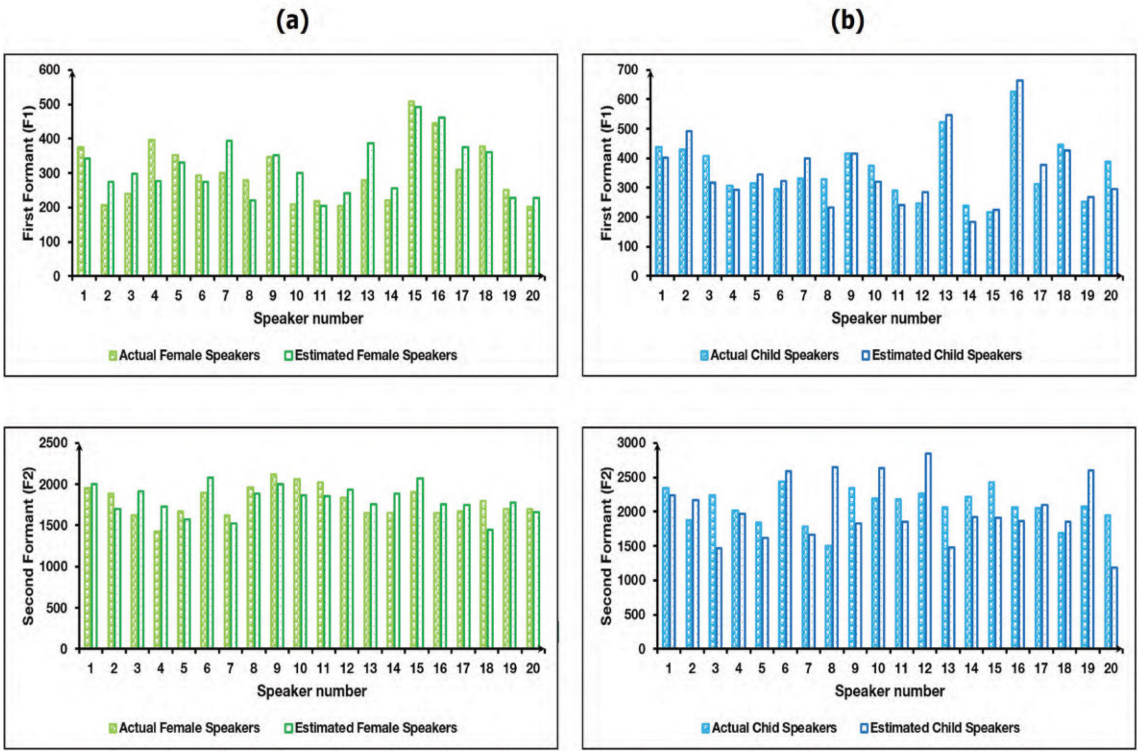


FIGURE 8. Graphical representation of the relationship between the formants of estimated and actual samples (a) female speakers (b) child speakers.

obtained from the relationship of males, females, and children were considered as estimated formants. We compared the actual and estimated formants of female and child speakers to validate the relationship. From Fig. 8, we observe that actual and estimated formants of female and child speakers are approximately close enough. Thus, we conclude that the relationship established between the male, female, and child formants are valid. Herein, the same relationship will be used for establishing the proposed vocal tract system for female and child speakers by using the Eq. (12).

The response of the complete proposed system was validated against the existing vocal tract system, which is discussed in Section III.

III. RESULTS AND DISCUSSION

The proposed vocal tract system was evaluated in terms of acoustic error (E) based on the frequency response and Normalized Root-Mean-Square Error (NRMSE) between the synthesized sound of both the systems using a Klatt synthesizer [48] or a speech synthesizer (speech synth) [49], against the existing system [5] for male, female and child speakers of English vowels and consonants as given in Eqs (13) and (14).

1) Acoustic error (E)
The acoustic error (E) [33] is defined as the mean square of the relative error (Mean Square Error) between both the systems, given by

$$E = 100\% * \sqrt{\frac{1}{2}((1-\frac{F_1}{\hat{F}_1})^2 + (1-\frac{F_2}{\hat{F}_2})^2} \qquad (13)$$

where F_1, F_2, and $\hat{F}_1$, $\hat{F}_2$ indicates the first two formants of the existing and proposed system, respectively. The proposed system responses are very close to existing system responses if the expected acoustic error is minimum.

2) Normalized Root-Mean-Square Error (NRMSE)
In statistical modeling and particularly regression analyses, a common way of measuring the quality of the fit of the model is the NRMSE (also called Root-Mean-Square Deviation) [50], given by

$$NRMSE = \sqrt{\frac{1}{2}\Sigma_{i=1}^{2}\left(\frac{F_i-\hat{F}_i}{F_i}\right)^2} \qquad (14)$$

where F_i and $\hat{F}_i$ are existing and proposed system formant values. The proposed system responses are very close to the existing system responses, if NRMSE will be minimum.

A. ACOUSTIC ERROR (E) BASED ON THE FORMANTS OF BOTH SYSTEMS

The frequency response of both existing and proposed systems for male, female, and child speakers for the vowels /a/, /e/, /i/, /o/, and /u/ and consonants /j/, /w/, /p/, /d/, /t/, /d/, /k/, /g/, /m/, and /n/ are shown in Figs. 9 and 10, respectively. There was a slight variation in the magnitude plot due to the cascade of the gain constants of systems compared with the existing system.

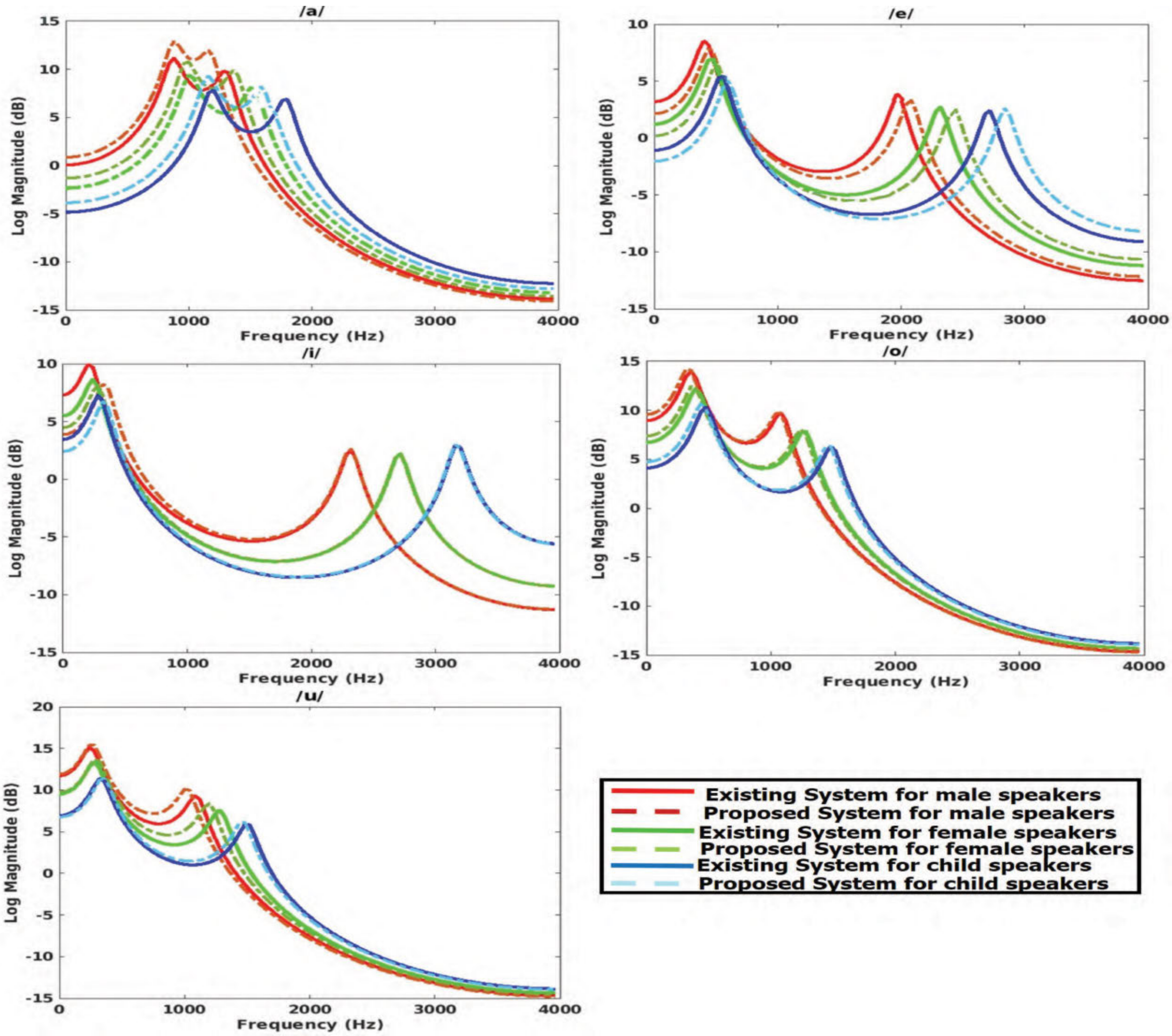


FIGURE 9. The frequency responses of the existing [5] and proposed vocal tract systems for the vowels using formants of male, female, and child speakers.

The acoustic error was calculated between the formants of both systems using Eq. (13) for male, female, and child speakers for vowels and consonants, as shown in Fig. 11.

The acoustic errors were comparatively higher for vowels /i/ and /u/ compared to /a/, /e/, and /o/ for the male, female, and children speakers, as shown in Fig. 11a. Similarly, acoustic errors were comparatively higher for consonants /j/, /w/, /k/, and /m/ than /p/, /b/, /t/, /d/, /g/, and /n/ for the male speakers. Similarly, acoustic error was slightly higher for /w/, /b/, and /k/ for female and /g/, /m/, and /j/ for child speakers compared with remaining consonants, as shown in Fig. 11b. These errors are due to assumptions of the articulatory gestures positions (tongue height, tongue advancement, burst, duration) in the formulation of tongue articulatory system formants, because of the limitations of the current measurement technology, as there is no precise data about the articulatory gestures.

The acoustic errors are relatively small at <5% for the proposed vocal tract system With the tongue articulatory system against the existing system. The formants of the proposed vocal tract system fall within the min-max range of the formant frequencies for the existing system [5] for male, female, and child speakers of the English alphabet.

B. NORMALIZED ROOT-MEAN-SQUARE ERROR BETWEEN THE SYNTHESIZED SOUND OF BOTH THE SYSTEMS USING FORMANT SYNTHESIZER

The synthesized sound waveforms generated using a formant synthesizer based on the vowel and consonants formants of the proposed system with tongue articulatory system against the existing system for male, female, and child speakers are shown in Figs. 12–16.

The NRMSE between the synthesized sound of both the systems of male, female, and child speakers for the English vowels and consonants is shown in Fig. 17, using Eq. (14).

The NRMSE was comparatively higher for /i/ and /u/ compared to /a/, /i/, and /o/ for the male speakers and comparatively higher for /o/ and /u/ compared to /a/, /e/, and /i/ for the

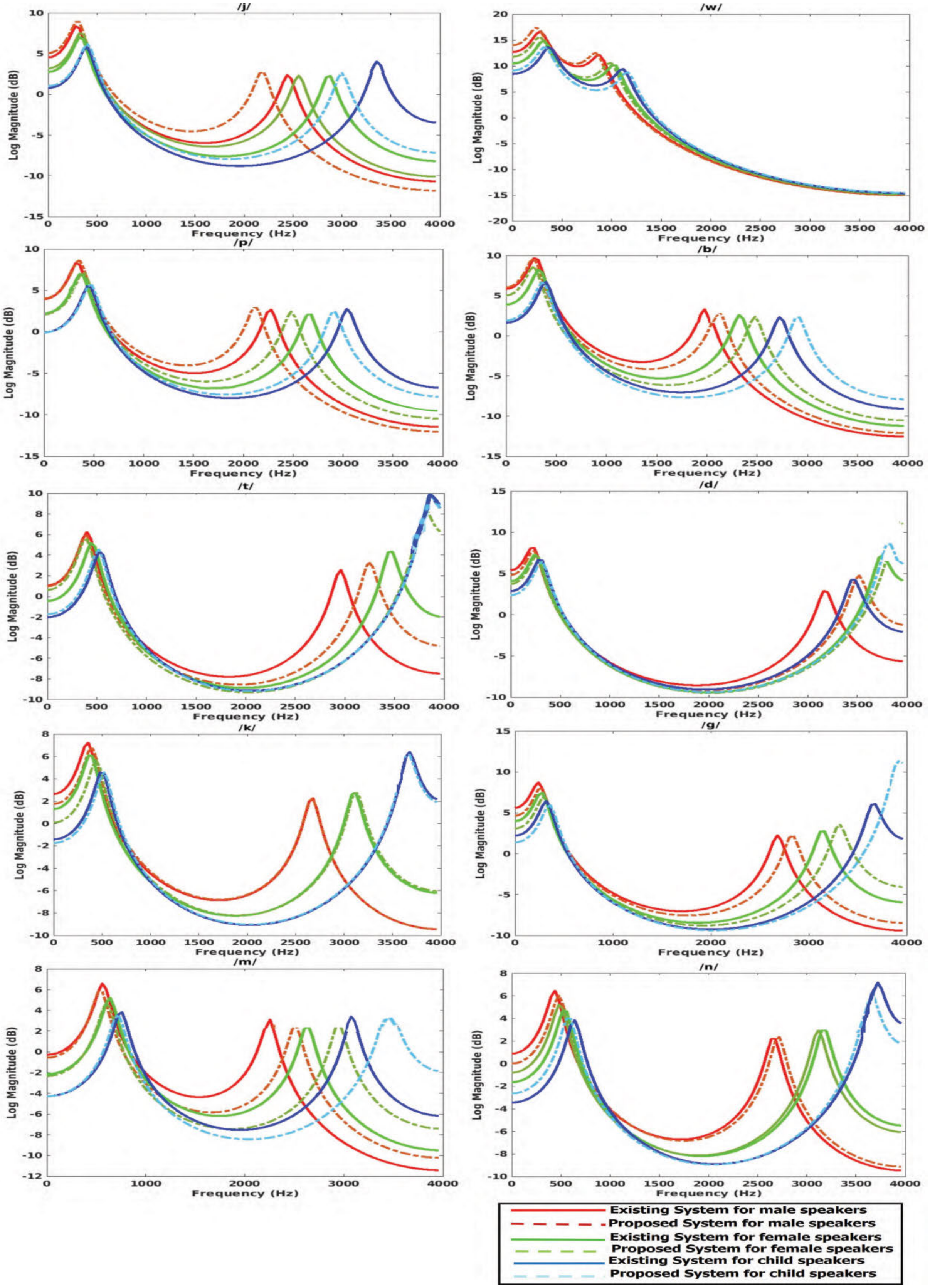


FIGURE 10. The frequency responses of the existing [5] and proposed vocal tract systems for the consonants using formants of male, female, and child speakers.

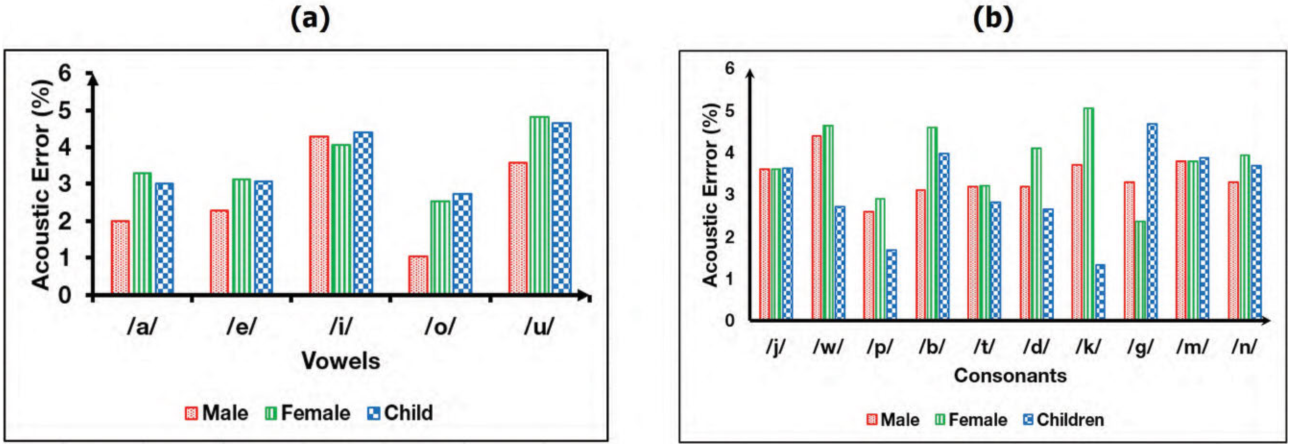


FIGURE 11. Acoustic errors between the existing and proposed systems for male, female, and child speakers (a) vowels and (b) consonants.

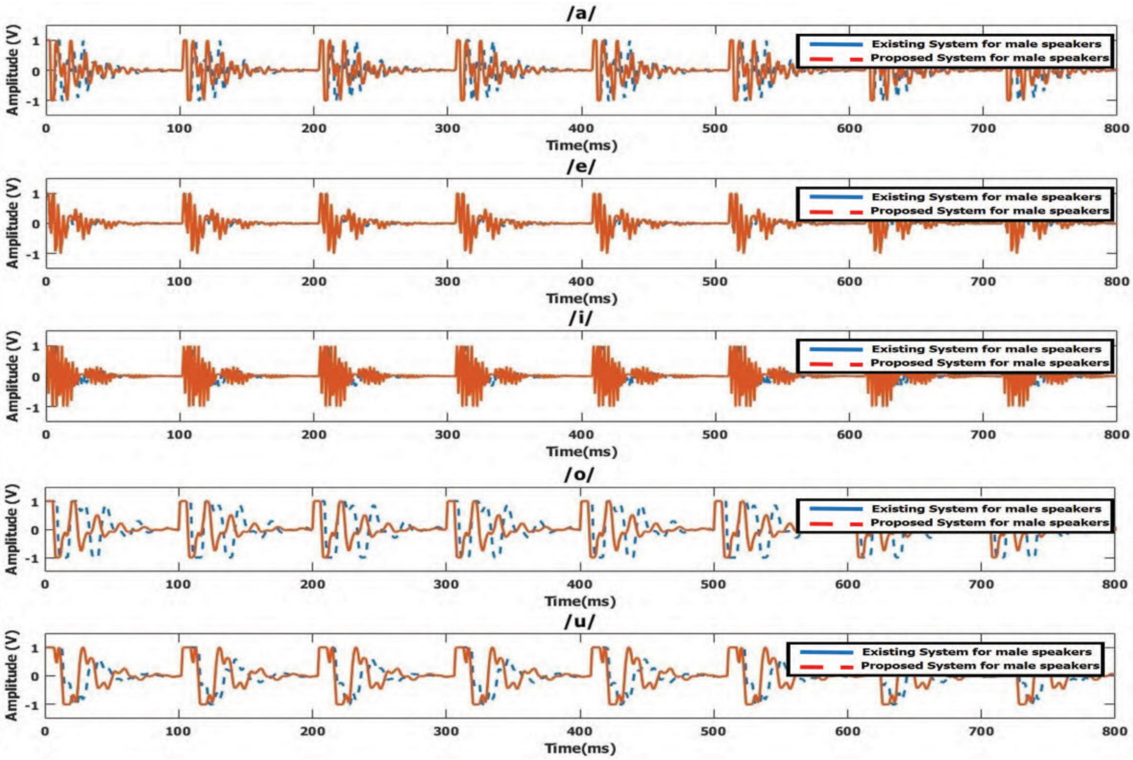


FIGURE 12. Synthesized speech waveform for existing and proposed systems of the English vowels using a formant synthesizer for male speakers.

female and child speakers, as shown in Fig. 17a. Similarly, NRMSE were comparatively higher for the consonants /j/, /w/, /k/, /m/ compared to /p/, /b/, /t/, /d/, /g/, and /n/ for the male speakers and comparatively higher for /t/, /g/, and /d/ compared to /j/, /w/, /p/, /b/, /k/, /m/, and /n/ for the female speakers and higher for /t/, /d/, /k/, and /n/ compared to /j/, /w/, /p/, /b/, /g/, and /m/ for the child speakers as shown in Fig. 17b. This difference highlights errors resulting from the simplifying assumptions made in developing the tongue articulatory system of the proposed vocal tract system.

By analyzing the formants between existing and proposed vocal tract systems in terms of acoustic error and NRMSE are with < 5% and < 0.15ms, respectively, for each English alphabet of male, female, and child speakers. Thus, the proposed vocal tract system is modeled and is close to the existing vocal tract system with negligible error.

Hence, a simple speech production system can be developed by using the tongue articulatory system especially based on tongue orientation characteristics.

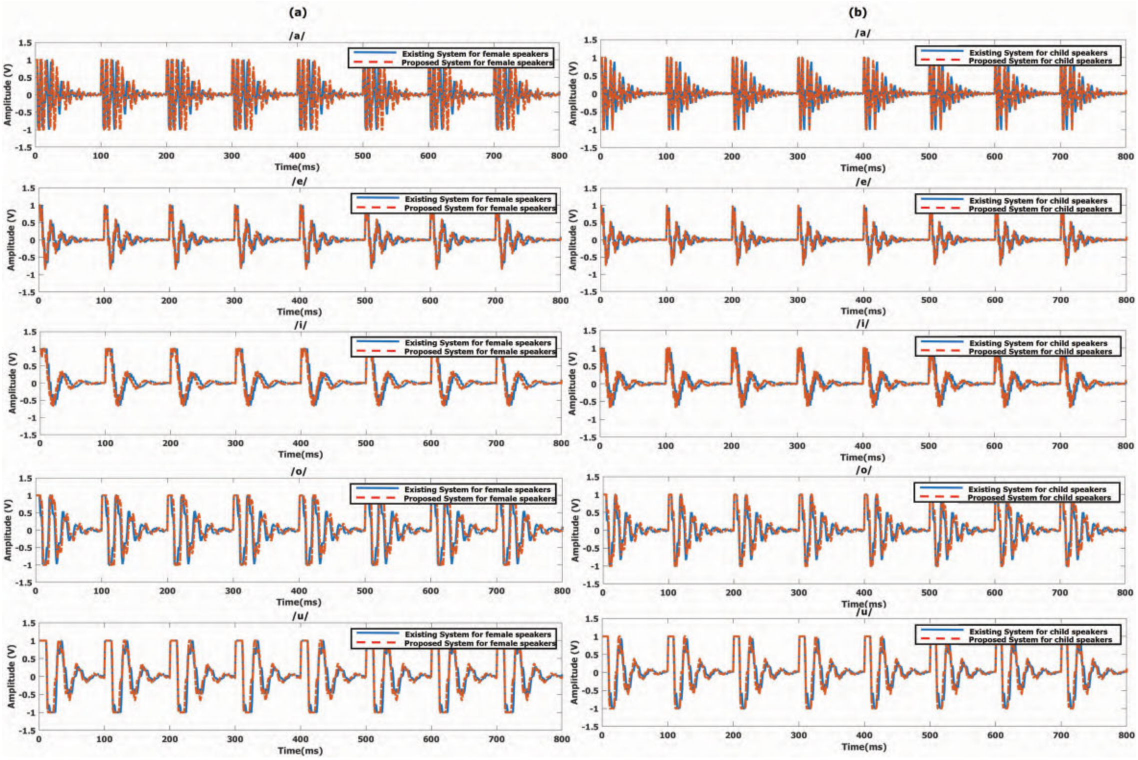


FIGURE 13. The synthesized sound waveform for existing and proposed systems of the English vowels for (a) female and (b) child speakers.

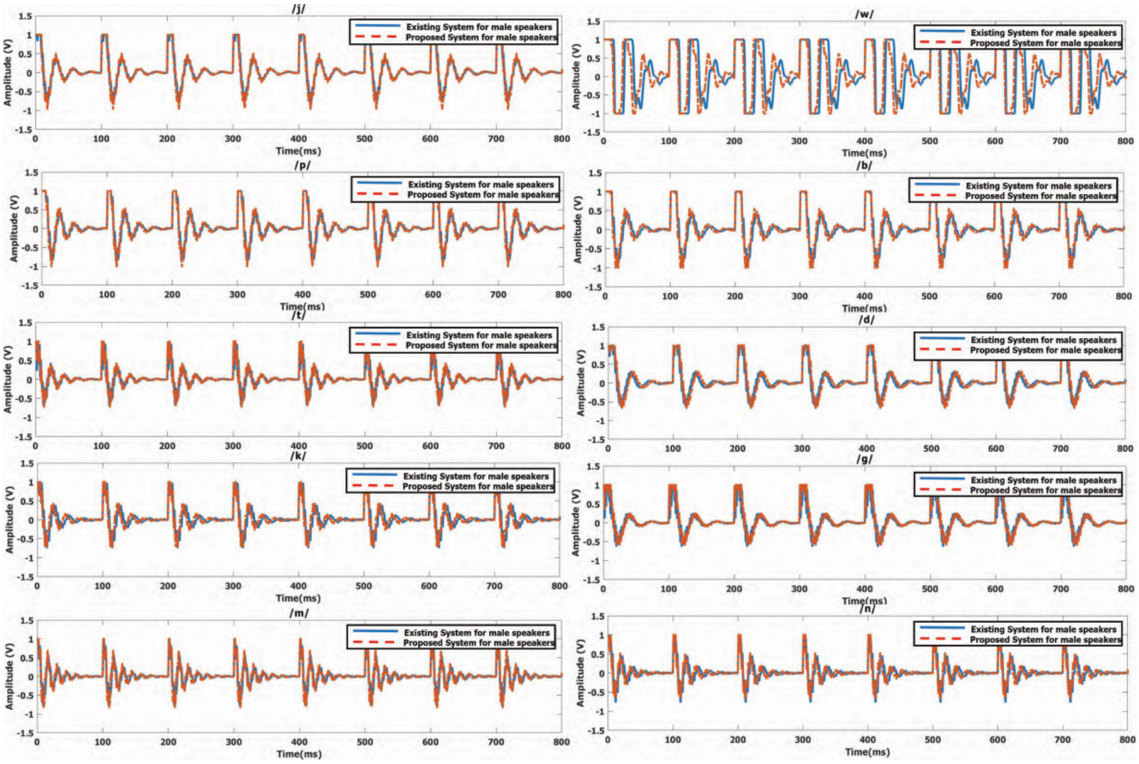


FIGURE 14. The synthesized sound waveform for existing and proposed systems of the English consonants for male speakers.

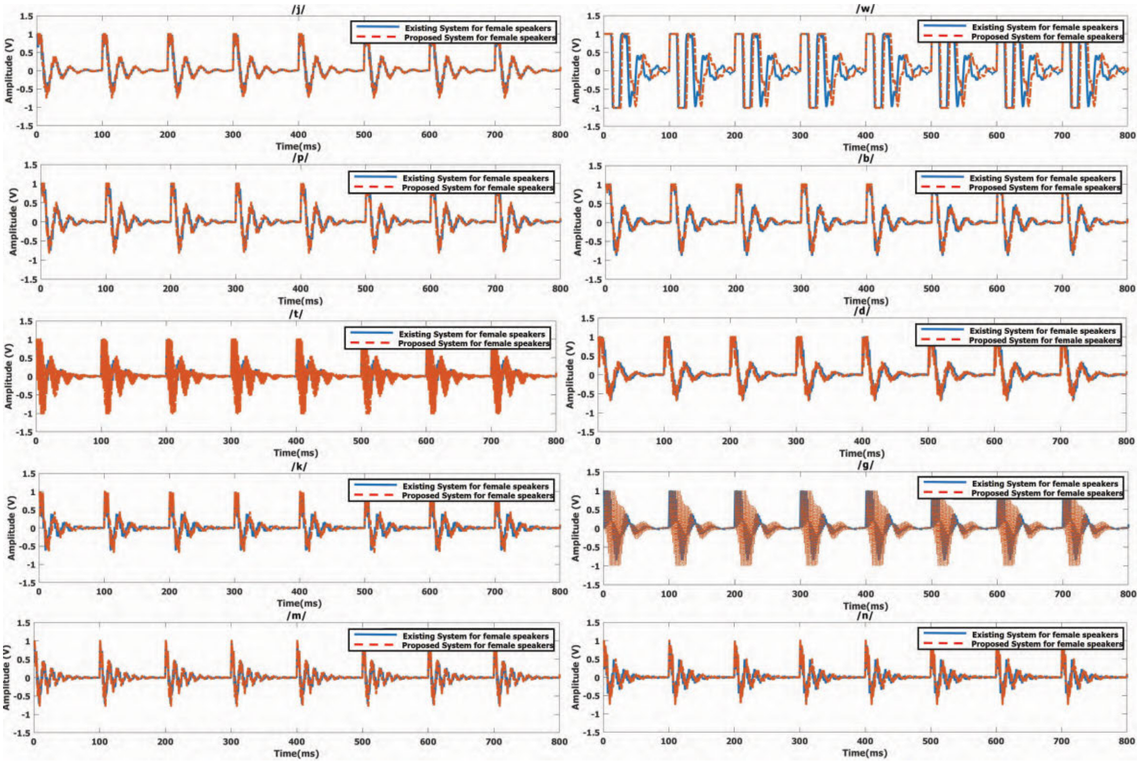


FIGURE 15. The synthesized sound waveform for existing and proposed systems of the English consonants for female speakers.

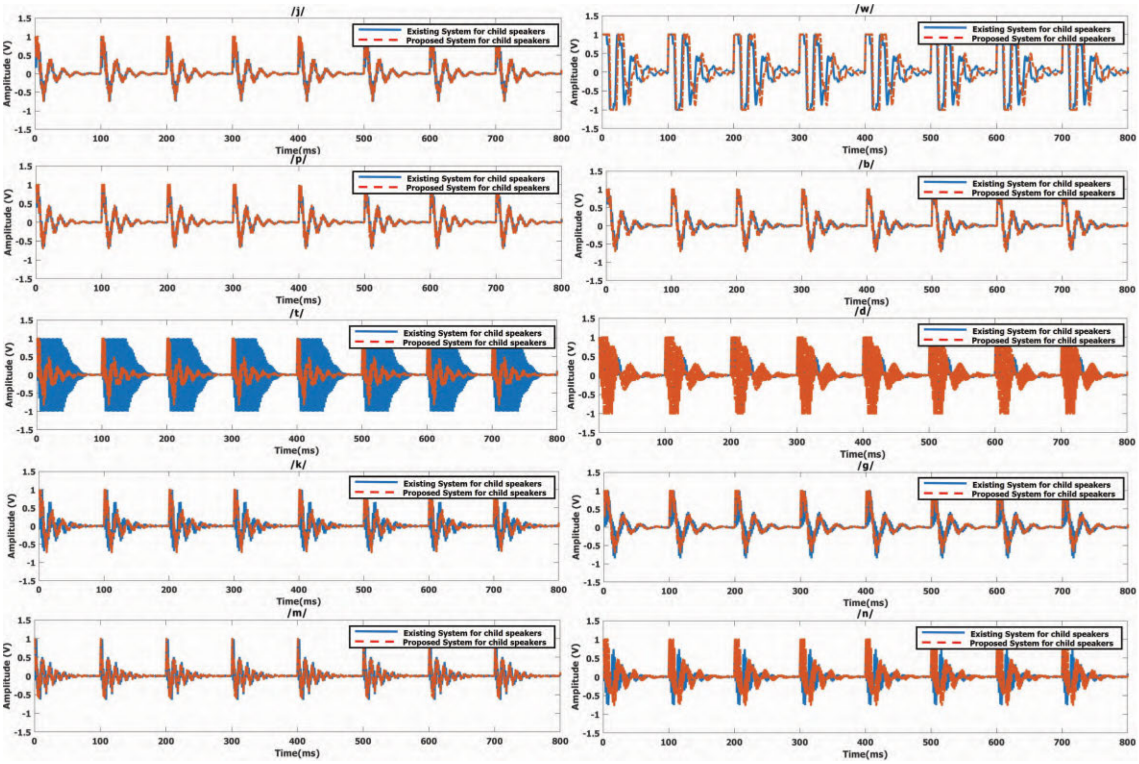


FIGURE 16. The synthesized sound waveform for existing and proposed systems of the English consonants for child speakers.

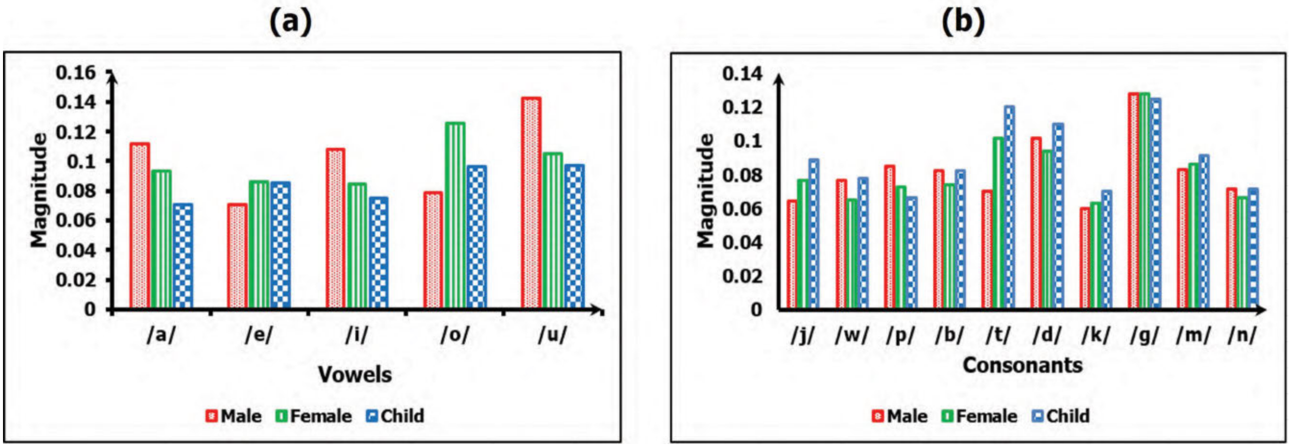


FIGURE 17. NRMSE for synthesized sound from existing and proposed systems for male, female, and child speakers: (a) vowels and (b) consonants.

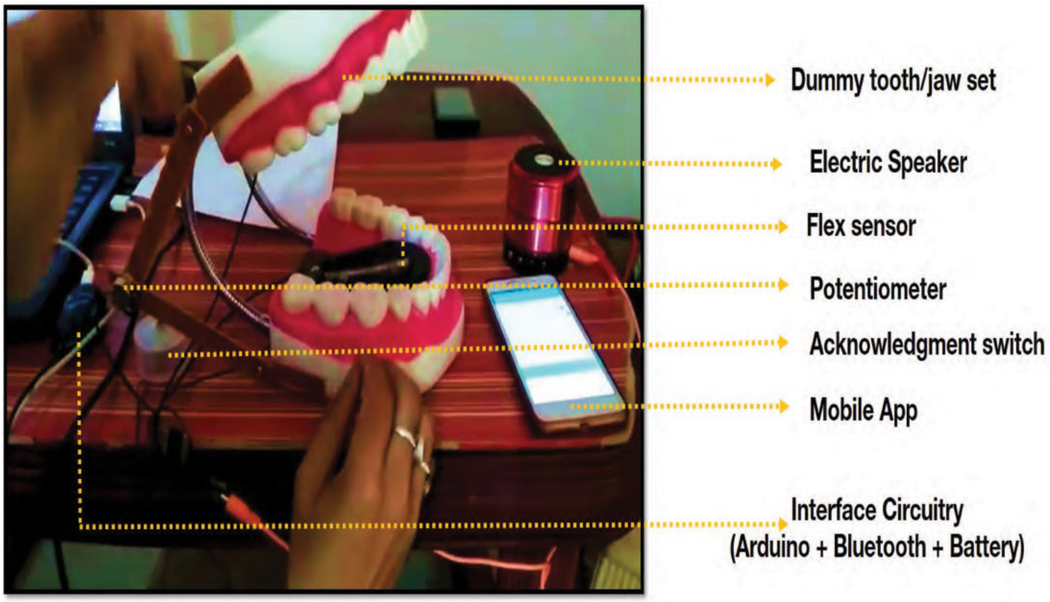


FIGURE 18. Hardware components of the proposed system for sound production.

IV. REAL-TIME SPEECH SYNTHESIS SYSTEM USING TONGUE MOVEMENTS

The proposed system was implemented in a dummy tooth model to produce English vowels and consonants using two important sensors as shown in Fig. 18. This section covers the proposed experimental hardware setup and proposed results and analysis and discussed briefly in the following subsections.

A. PROPOSED HARDWARE EXPERIMENTAL SETUP

The artificial tooth set was used as a hardware prototype, with sensors representing jaw and tongue movement flexibly. Herein, we considered two important sensors, namely, the flex sensor and potentiometer, for tracking the position of the tongue specifically tongue height and tongue advancement, which was given as input to the dummy tooth model. The sensor values were displayed on a laptop screen through Arduino and on a mobile screen through a serial monitor application. The output sound can hear from an electric speaker. Mini-prototype hardware setup was used with rechargeable battery power for sound production, as shown in Fig. 18.

There are two independent aspects to the hardware, namely dummy tooth/jaw set and interface circuitry.

1) Dummy tooth/jaw set: A dummy tooth/jaw set is used by dentists to show the teeth, gums, and cavity to patients. It is used as a dummy human tooth/jaw model. It is not possible to fix sensors inside the mouth of a person since sensors are at a macroscale. There are two sensors (flex, potentiometer) affixed to the dummy

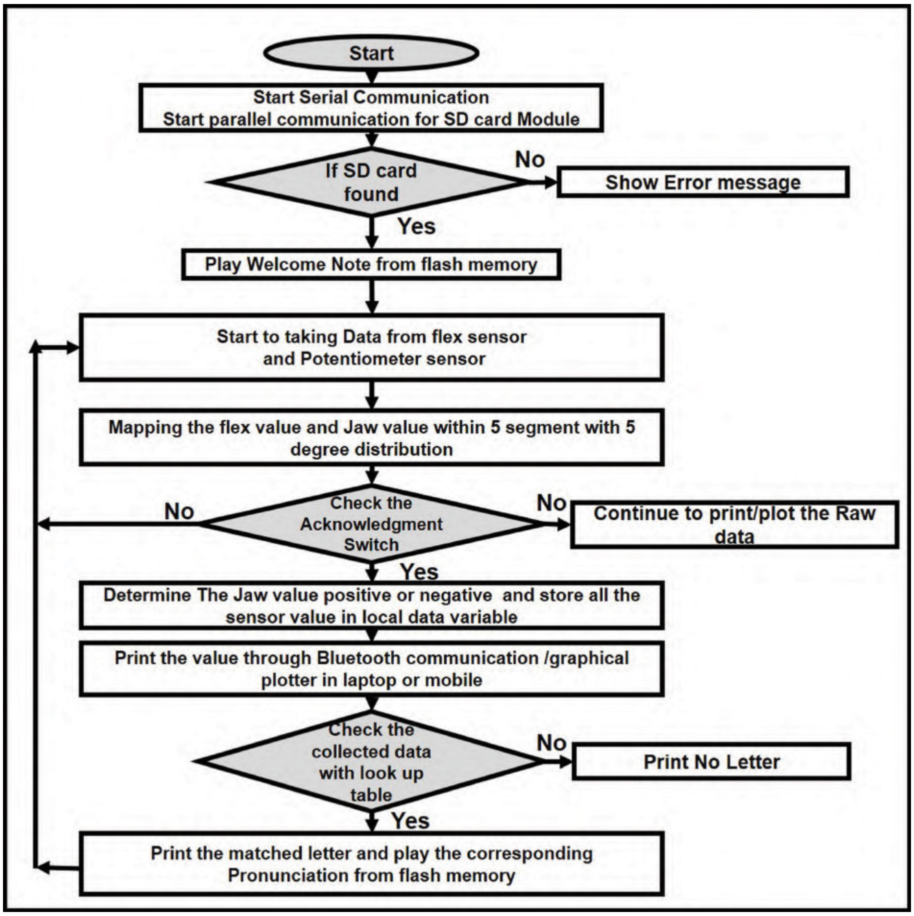

Start
Start Serial Communication
Start parallel communication for SD card Module
If SD card found
No
Show Error message
Yes
Play Welcome Note from flash memory
Start to taking Data from flex sensor and Potentiometer sensor
Mapping the flex value and Jaw value within 5 segment with 5 degree distribution
No
Check the Acknowledgment Switch
No
Continue to print/plot the Raw data
Yes
Determine The Jaw value positive or negative and store all the sensor value in local data variable
Print the value through Bluetooth communication /graphical plotter in laptop or mobile
Check the collected data with look up table
No
Print No Letter
Yes
Print the matched letter and play the corresponding Pronunciation from flash memory

FIGURE 19. Flow diagram of the proposed hardware setup.

TABLE 4. The hardware components required for the experimental setup.

S.No	Hardware component	Functional	Usage
Dummy tooth/jaw set			
1	Flex sensor	Used to capture the movement of the tongue and the tongue tip, set for analog value (0-5). 0 define there is no movement of the tongue and 5 defines tongue tip touches the upper palate for certain alphabets.	Tongue movement involves bending and rolling, so the closest sensor that can replicate the movement of tongue is flex sensor.
2	Potentiometer	Pronunciation of alphabets requires movement of upper and lower jaw as well. since the user always keep the dummy tooth/jaw set on a table, the lower jaw is almost always immovable, to add benefit to the user. Thus, the upper jaw moves the value is registered as (0 -5).	The jaw movement during pronunciation of each alphabet is different. The best way to capture the various movements of the jaw is via variable resistor that gives a variable output voltage.
Interface circuitry			
3	Arduino Nano V3.0	Used for programming the Bluetooth module-HC05 for transmission of data acquired by sensors from dummy tooth/jaw set.	It is very small, compact and has all the functionalities of Arduino UNO.
4	Bluetooth module - HC05	Communicate wirelessly from the dummy tooth/jaw set to the speech assistant device via this module.	The user will face difficulties connecting the dummy tooth/jaw set to the speech assistant device by wires.It will give better mobility and portability options for the user if it is wireless, so bluetooth is the best option.

tooth/jaw set to acquire data/values that are required to pronounce a particular English alphabet.

2) Interface circuitry: Interface circuitry enables the communication between a dummy tooth/jaw set and speech assistant device: electric speaker and mobile application. Whatever values the sensors acquire from the various user-defined movements of the dummy tooth/jaw set can be sent to the speech assistant device.

The hardware components required for the hardware setup are enumerated in Table 4.

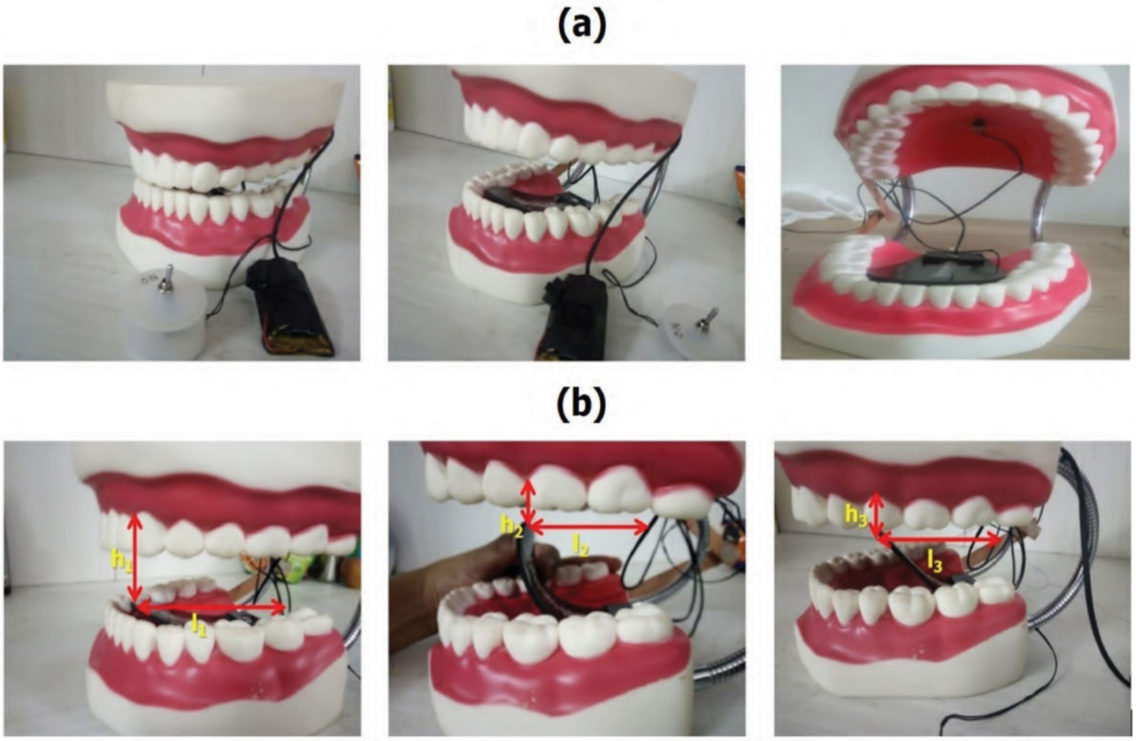


FIGURE 20. (a) Jaw at different positions i.e. mouth closing and opening gestures (b) Various tongue height and advancement gestures.

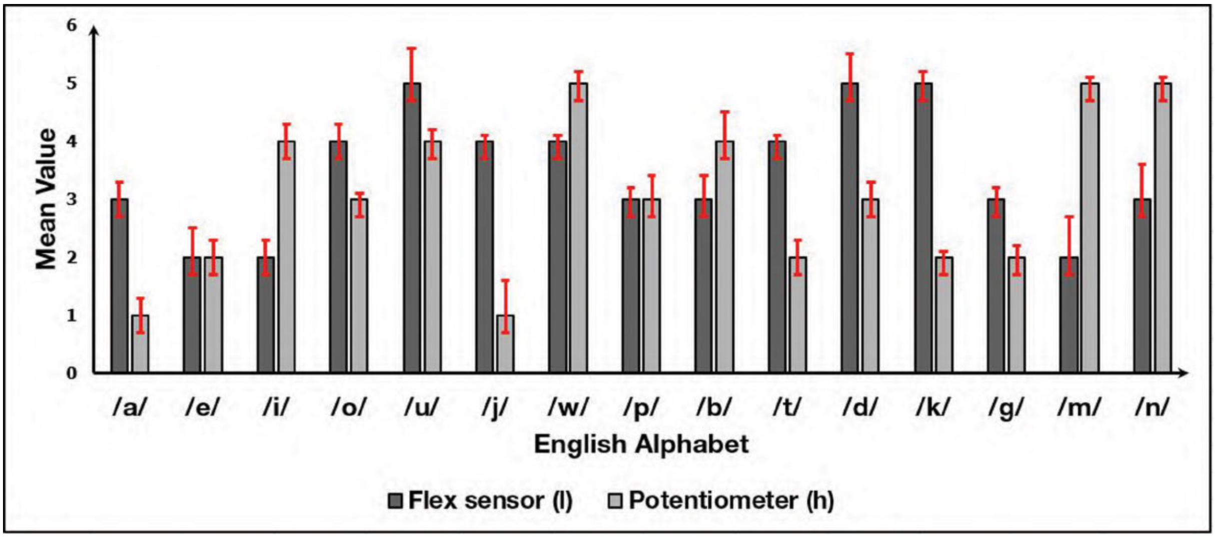
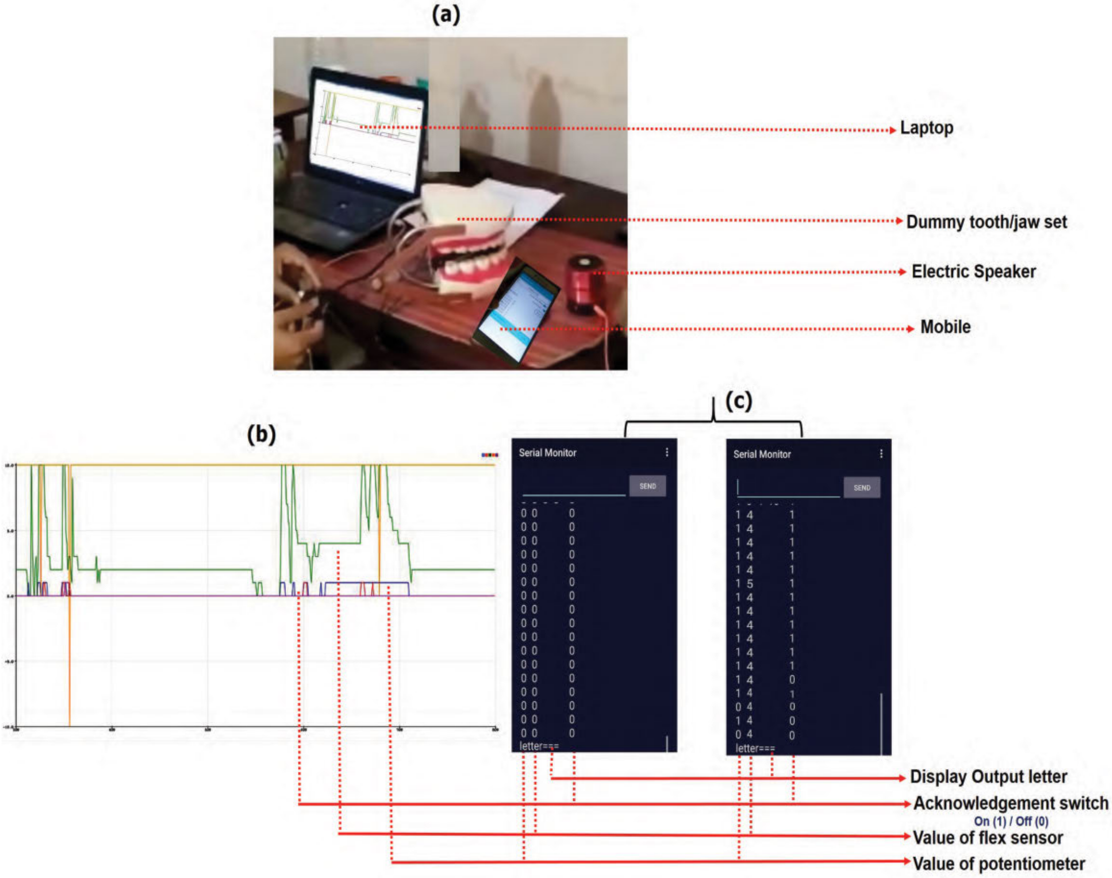


FIGURE 21. The value of flex sensor (*l*) and potentiometer (*h*) in centimeters for the English alphabet.

An algorithm was developed to pronounce every English alphabet distinctively. The rolling of the tongue in multiple degrees and subsequent touch was observed through sensors and recorded. The look-Up table (LUT) was designed in an optimized way to be coded in the microcontroller's memory. The pronounced alphabet was heard through the electric speaker using parallel communication with a micro SD card. Some formative training is necessary to use it. The algorithm steps used the hardware setup of a real-time speech synthesis system using tongue movements is shown in Fig. 19.

Flex sensors provided the position of the tongue, i.e., the tongue advancement value, whereas a jaw sensor called a potentiometer provided the tongue height value. Depending on the movements of oral cavity gestures, we started to



```
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
0 0     0
letter===
```

```
1 4     1
1 4     1
1 4     1
1 4     1
1 4     1
1 5     1
1 4     1
1 4     1
1 4     1
1 4     1
1 4     1
1 4     1
1 4     0
1 4     1
0 4     0
1 4     0
0 4     0
letter===
```

FIGURE 22. Displays value of acknowledgment switch, flex and potentiometer in (a) proposed hardware setup (b) laptop screen (c) mobile screen.

TABLE 5. Expected and wrongly produced output sounds produced from the proposed hardware setup.

English alphabet	Expected produced output sound based on sensor value	Wrongly produced output sound based on sensor value beyond error margin (see figure 21)
Vowels	/a/	/o/
	/e/	No letter
	/i/	No letter
	/o/	/u/
	/u/	/o/
Consonants	/j/	/g/
	/w/	No letter
	/p/	No letter
	/b/	/n/
	/t/	/g/
	/d/	/p/
	/k/	/d/
	/g/	/t/
	/m/	No letter

obtain data from the flex and jaw sensor, which was given as input to a dummy tooth model. The oral cavity gestures at different positions during articulation of different sounds are shown in Fig. 20. The optimized way of LUT used in the hardware setup, contains the mean value of flex (l) and potentiometer (h) sensors, whose calibrated values are measured in centimeters, are shown in Fig. 21 for English speech production. Error bars represent the standard error of the mean. The X-axis represents the respective English vowel or consonant. Y-axis represents the reference mean value of tongue height (h) and tongue advancement (l) obtained using flex and potentiometer sensors. Fig. 21 shows, a neat visualization of reference mean value of sensors with respective of English vowels and consonants. The system checks whether the acknowledgment switch is ON. Mapping the flex and jaw value store in the local data variable gets printed through Bluetooth communication/graphical plotter of Arduino on a laptop screen and in mobile through serial monitor application. The collected sensor data was checked using a LUT (see Figure 21). If sensor data did not match with the LUT, NO letter was printed on the screen, and a corresponding sound was pronounced. If there was a match of sensor data with the LUT, then the function called for audio play through serial-parallel interface communication with SD card (flash memory), the letter sounds were stored previously through the electric speaker and the relevant letter was printed on the laptop and mobile screen. The steps were repeated while taking the sensor data continuously.

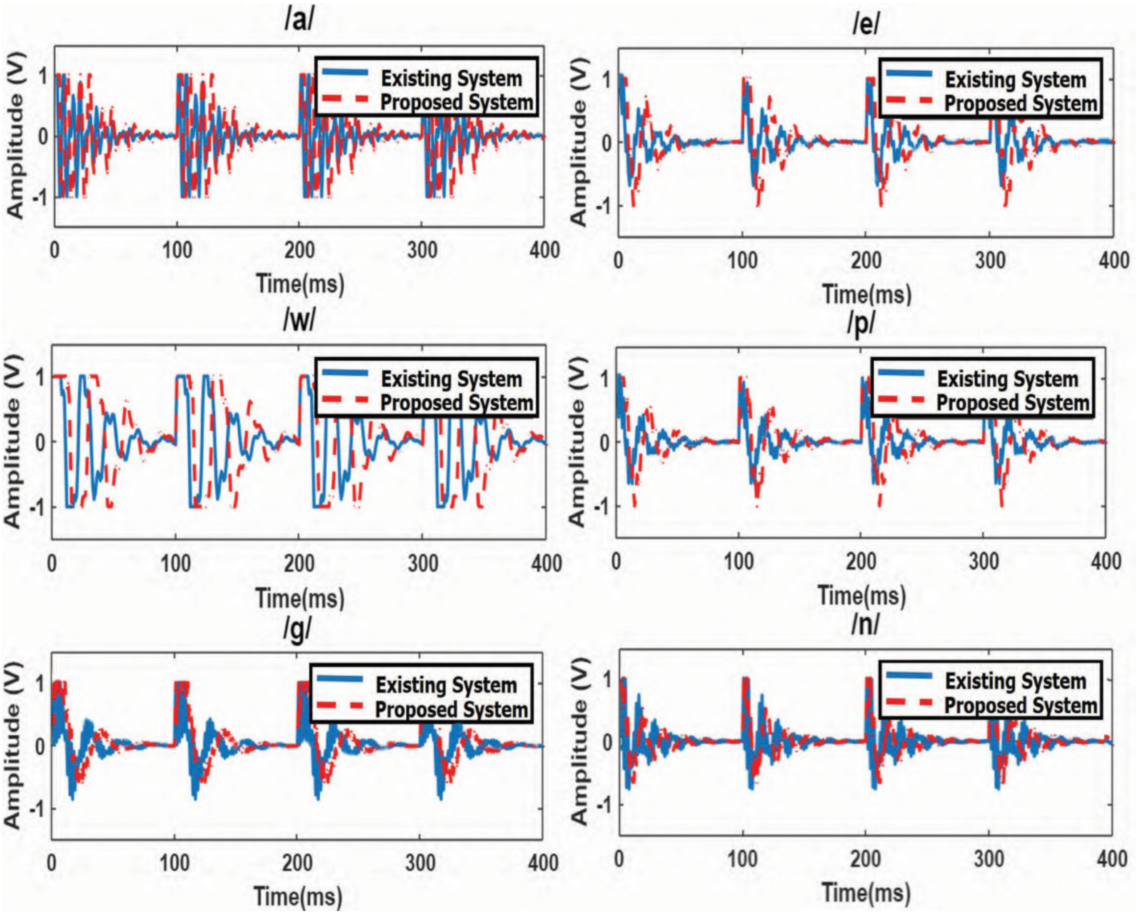


FIGURE 23. The output waveform of respective sound produced from the existing and proposed hardware system.

B. EXPERIMENTAL SETUP OUTPUT RESULTS AND ANALYSES

The proposed hardware setup for speech analyses with laptop and mobile is shown in Fig. 22a. The variability of flex and potentiometer sensor values were displayed on a laptop screen through Arduino and on mobile through the serial controller, as shown in Fig. 22b and 22c, respectively. The respective output sound was heard from the electric speaker from the letter sounds previously stored on the SD card based on the matching of sensor data with the LUT.

When the acknowledgment switch was On {1}, the hardware setup read the sensor data. After that, when the acknowledgment switch was Off {0}, it displayed the relevant English letter based on the match of sensor data with the LUT, and the same letter was pronounced through the electric speaker from the sounds stored previously on the SD card, whose output sound waveforms are shown in Fig. 23 against an existing system (human speech production).

Beyond a certain range of values in the combinations of flex (l) and potentiometer (h) sensors, which are indicated in the LUT (see Fig. 21), the system may lead to producing NO or a different alphabet than intended. We further analyzed tolerance limits based on the variability impact of the sensor on the experimental setup for sound production. Sometimes, due to electrical stress on sensors, the combination of input may vary, as it is difficult to set the particular sensor value, resulting in pronouncing the wrong sound. A wrongly produced output sound was considered heard based on the sensor value given as an input to the proposed hardware setup beyond the error margin (see Fig. 21) instead of the expected output sound. A list of expected and wrongly produced output sounds were obtained from a tolerance analysis of 'l' and 'h' combinations, respectively, are provided in Table 5 followed by the related waveforms in Fig. 24.

The original sounds expected based on the sensor input are /a/, /o/, /e/, /j/, /g/, and /w/ but the electrical stress on the sensor (sensor value +/− tolerance value) may produce wrong sounds, such as /o/, /u/ and no sound, /g/, /t/ and no sound, as shown in Fig. 24.

The perceptual test [51] is validated by a total of 30 human subjects of both sexes, aged between 11 and 52 years with normal hearing and were normally developed. A perception score test was determined to evaluate the quality of output speech produced from the proposed hardware setup. The perception scores measure the extent to which listeners misjudge the sound. The ranking scale used for the perception

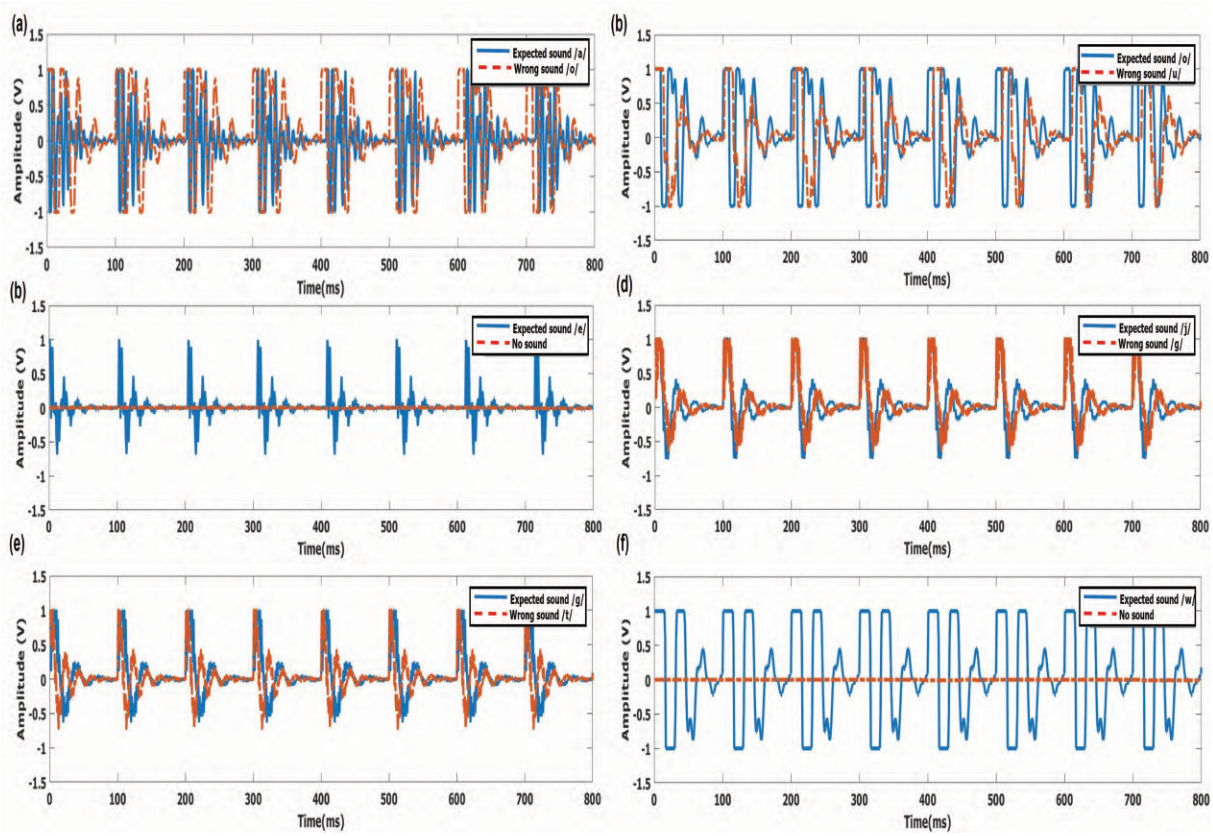


FIGURE 24. The output waveform of expected and wrongly produced sounds from the proposed hardware setup.

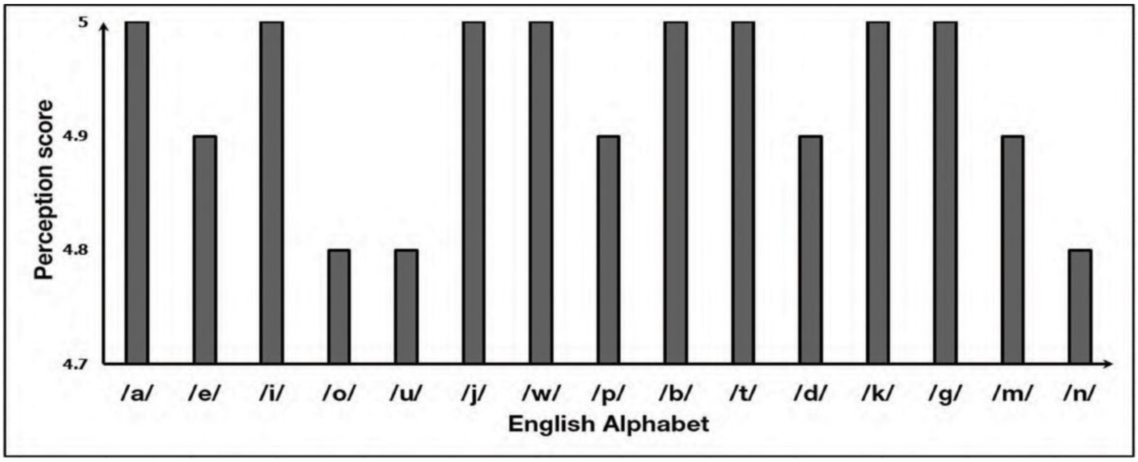


FIGURE 25. Results of the perception test averaged over listeners for each English alphabet.

score was 1–5, with five being the best score and one the worst score. The score during the perception test is provided in Fig. 25, averaged over listeners for each English alphabet. The mean perception scores by listeners for vowels and consonant groups are given in Table 6.

The overall performance in speech perception ability for all the age groups of listeners is quite good as shown in Fig. 25. The low perception scores were observed particularly for vowels /o/ and /u/ and significantly low perception scores for consonants /p/, /d/, /m/ and /n/ because of noisy environments or due to their limited vocabulary and language skills. The individual perception scores for all vowel and consonant groups by all the listeners are summarized and listed in Table 6, in which vowels and nasals groups have relatively low perception scores. Thus, the overall output performance of the proposed hardware system was equivalent

TABLE 6. Mean perception score during the perceptual test.

English alphabet (groups)	Mean perception score during perceptual test
Vowels	4.97
Consonants	
Approximants	5
Plosives	4.96
Nasals	4.85

to natural speech based on the performance of initial English vowels and consonants.

V. CONCLUSION

The formant frequencies of the oral cavity system based on tongue orientation characteristics for vowel and consonant sound production was estimated using optimized statistical relation. We compared the proposed system with an existing system to validate the estimated formant frequencies of the tongue articulatory system. The responses obtained for the existing and proposed systems were similar. The outputs were validated quantitatively using acoustic error based on the formants from frequency response and Normalized Root-Mean-Square Error (NRMSE) from synthesized sound using a formant synthesizer. The proposed system was comparable to the existing system with an acoustic error for all English vowels and consonants of male, female, and children speakers at approximately $< 5\%$. All formant synthesized outputs from the existing and proposed systems were compared with values of approximately < 0.15ms, and the perceptive analyses offered satisfactory results.

The theoretical and practical analyses presented here can be used for developing replacements of the glottis and laryngeal system through a sequence of amplified pulses with a specific frequency, i.e., pitch. The formants of the oral cavity system for vowels and consonants based on tongue gestures, bypassing the glottal portion can then be used to generate speech for the speech-disabled. Hence, the hardware system is designed and demonstrated using an experimental dummy tongue model setup with sensors analyzed for speech production. The output sounds are heard from the electric speaker and displayed on the same in a laptop through Arduino and a mobile screen using a serial monitor application, which is applied to complete speech production for patients with speech disorders.

We intend to extend the work to synthesize words and sentences based on the degree of articulators to enable easy communication for the speech disabled. In future enhancement, this idea will extend to develop a wearable device that contains saliva proof sensors placed on the appropriate removable fixture, with a miniature custom-built electric speaker or one placed near the mouth either in a Bluetooth module or neck to produce real-time speech.

VI. ACKNOWLEDGMENT

The authors would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India, for financial support during the research.

REFERENCES

[1] K. Bunning, J. K. Gona, V. Odera-Mung'ala, C. R. Newton, J.-A. Geere, C. S. Hong, and S. Hartley, "Survey of rehabilitation support for children 0–15 years in a rural part of Kenya," *Disab. Rehabil.*, vol. 36, no. 12, pp. 1033–1041, Jun. 2014.

[2] 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*. Hyattsville, MD, USA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Jun. 2015.

[3] 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," *Int. J. Lang. Commun. Disorders*, vol. 35, no. 2, pp. 165–188, Apr. 2000.

[4] L. D. Shriberg, J. B. Tomblin, and J. L. McSweeny, "Prevalence of speech delay in 6-year-old children and comorbidity with language impairment," *J. Speech, Lang., Hearing Res.*, vol. 42, no. 6, pp. 1461–1481, Dec. 1999.

[5] L. Rabiner, "Fundamentals of speech recognition," in *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[6] 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 Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 325–332, Feb. 2004.

[7] J. Wang, A. Samal, and J. R. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," in *Proc. 5th Workshop Speech Lang. Process. Assistive Technol. (SLPAT)*. Baltimore, MD, USA: Association for Computational Linguistics, Aug. 2014, pp. 38–45.

[8] A. Katsamanis, E. Bresch, V. Ramanarayanan, and S. Narayanan, "Validating rt-MRI based articulatory representations via articulatory recognition," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 28–31.

[9] 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," *Open Rehabil. J.*, vol. 3, no. 1, pp. 1–7, Apr. 2010.

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

[11] A. S. Dilbagi. *16-Year-Old Invents A Breath Enabled Talking Device to Help the Speech Impaired*. Accessed: 2014. [Online]. Available: http://www.thebetterindia.com

[12] H. Sahni, A. Bedri, G. Reyes, P. Thukral, Z. Guo, T. Starner, and M. Ghovanloo, "The tongue and ear interface: A wearable system for silent speech recognition," in *Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2014, pp. 47–54.

[13] K. A. U. Menon, R. Jayaram, and P. Divya, "Wearable wireless tongue controlled assistive device using optical sensors," in *Proc. 10th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Jul. 2013, pp. 1–5.

[14] K. A. U. Menon, R. Jayaram, D. Pullarkatt, and M. V. Ramesh, "Wearable wireless tongue controlled devices," U.S. Patent 9 996 168, Jun. 12, 2018.

[15] C. Sokolowski, "The tongue: Vowel formation," Ph.D. dissertation, School Music Partial Fulfillment Requirements Degree Master Music, Indiana Univ., Bloomington, IN, USA, 2014.

[16] P. Padmini, S. Tripathi, and K. Bhowmick, "Sensor based speech production system without use of glottis," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 2073–2079.

[17] G. R. Divya, H. Jayamohanan, N. V. Smitha, R. Anoop, A. Nambiar, T. Krishnakumar, and K. Pavithran, "Primary neuroendocrine carcinoma of the larynx: A case report," *Indian J. Otolaryngol. Head Neck Surgery*, Aug. 2020, pp. 1–4, doi: 10.1007/s12070-020-02060-z.

[18] A. J. Venker-van Haagen, "Diseases of the larynx," *Vet. Clinics North Amer., Small Animal Pract.*, vol. 22, no. 5, pp. 1155–1172, Sep. 1992.

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

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

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

[22] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model," *Folia Phoniatrica et Logopaedica*, vol. 64, no. 2, pp. 64–72, 2012.

[23] B. E. F. Lindblom and J. E. F. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," *J. Acoust. Soc. Amer.*, vol. 50, no. 4B, pp. 1166–1179, Apr. 1971.

[24] S.-H. Lee, J.-F. Yu, Y.-H. Hsieh, and G.-S. Lee, "Relationships between formant frequencies of sustained vowels and tongue contours measured by ultrasonography," *Amer. J. Speech-Lang. Pathol.*, vol. 24, no. 4, pp. 739–749, Nov. 2015.

[25] J. Lee, S. Shaiman, and G. Weismer, "Relationship between tongue positions and formant frequencies in female speakers," *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 426–440, Jan. 2016.

[26] C. Johansson, J. Sundberg, and H. Wilbrand, "X-ray study of articulation and formant frequencies in two female singers," in *Proc SMAC*, 1985, vol. 83, no. 1, pp. 203–218.

[27] L. J. Raphael, F. Bell-Berti, R. Collier, and T. Baer, "Tongue position in rounded and unrounded front vowel pairs," *Lang. Speech*, vol. 22, no. 1, pp. 37–48, Jan. 1979.

[28] P. Ladefoged and R. Harshman, "Formant frequencies and movements of the tongue," in *Proc. UCLA Work. Papers Phonetics*, vol. 45, 1979, pp. 39–52.

[29] S. S. Narayanan, A. A. Alwan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," *J. Acoust. Soc. Amer.*, vol. 101, no. 2, pp. 1064–1077, Feb. 1997.

[30] M. Y. Chen, "Acoustic correlates of English and French nasalized vowels," *J. Acoust. Soc. Amer.*, vol. 102, no. 4, pp. 2360–2370, Oct. 1997.

[31] K. Nyman, "Cues to vowels in the aperiodic phase of English plosive onsets," Ph.D. dissertation, Dept. Lang. Linguistic Sci., Univ. York, York, U.K., 2011.

[32] J. G. Švec, J. Horáček, F. Šram, and J. Veselý, "Resonance properties of the vocal folds: *In vivo* laryngoscopic investigation of the externally excited laryngeal vibrations," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1397–1407, Oct. 2000.

[33] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, Apr. 2013, Art. no. e60603.

[34] A. Zourmand, S. Mirhassani, H.-N. Ting, S. Bux, K. Ng, M. Bilgen, and M. Jalaludin, "A magnetic resonance imaging study on the articulatory and acoustic speech parameters of malay vowels," *Biomed. Eng. OnLine*, vol. 13, no. 1, p. 103, 2014.

[35] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. LDC93S1, 1993.

[36] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area," *J. Acoust. Soc. Amer.*, vol. 134, no. 5, pp. EL477–EL483, Nov. 2013.

[37] S. Sitole. (2020). *Gradient Descent (Solving Quadratic Equations with Two Variables)*. MATLAB Central File Exchange. Accessed: Oct. 12, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/62010-gradient-descent-solving-quadratic-equations-with-two-variables

[38] D. Maurer *Acoustics of the Vowel-Preliminaries*. Bern, Switzerland: Peter Lang, 2016.

[39] K. N. Stevens, *Acoustic Phonetics*, vol. 30. Cambridge, MA, USA: MIT Press, 2000.

[40] M. G. Di Benedetto and A. Esposito, "Acoustic analysis and perception of classes of sounds (vowels and consonants)," in *Proc. Speech Process., Recognit. Artif. Neural Netw.* London, U.K.: Springer, 1999, pp. 54-84.

[41] M. Shariq, "Arabic and English consonants: A phonetic and phonological investigation," *Adv. Lang. Literary Stud.*, vol. 6, no. 6, pp. 146–152, Dec. 2015.

[42] S. Fuchs and P. Birkholz, "Phonetics of consonants," in *Oxford Research Encyclopedia of Linguistics*. Oxford, U.K.: Oxford Univ. Press, Jul. 2019.

[43] G. Fant, *Acoustic Theory of Speech Production*. Hague, The Netherlands: Mouton, 1960.

[44] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. London, U.K.: Pearson, 2006.

[45] L. Ménard, J.-L. Schwartz, L.-J. Boë, S. Kandel, and N. Vallée, "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1892–1905, Apr. 2002.

[46] J. Sundberg, "The KTH synthesis of singing," *Adv. Cognit. Psychol.*, vol. 2, no. 2, pp. 131–143, Jan. 2006.

[47] M. P. Gelfer and Q. E. Bennett, "Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender," *J. Voice*, vol. 27, no. 5, pp. 556–566, Sep. 2013.

[48] J. C. Rutledge, K. E. Cummings, D. A. Lambert, and M. A. Clements, "Synthesizing styled speech using the Klatt synthesizer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1995, pp. 648–651.

[49] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: Improving production quality," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1870–1887, Dec. 1989.

[50] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdiscipl. J. Inf., Knowl. Manage.*, vol. 14, pp. 45–76, Jan. 2019.

[51] V. B. Queiroz, N. E. Zamberlan-Amorim, K. J. Pinotti, E. A. da Silva Lizzi, and A. C. M. B. Reis, "Speech perception test with pictures: Applicability in children with hearing impairment," in *Proc. Rev. CEFAC*, Mar. 2017, pp. 180–189.

PALLI PADMINI received the B.Tech. degree in electronics and communications engineering from the Siddharth Institute of Engineering and Technology, Andhra Pradesh, India, in 2012, and the M.Tech. degree in digital electronics and communication systems from the KSRM College of Engineering, Kadapa, in 2014. She has a teaching experience of a year with the Siddharth Educational Academy Group of Institutions, Tirupati, India. She has conducted and participated in a number of short-term courses, seminars, and conferences conducted at the national level. She has been pursuing research with the Amrita School of Engineering, Bengaluru, India, since April 2016. She is also a full-time Ph.D. Scholar under the Ministry of Electronics and Information Technology, Government of India's Visvesvaraya Ph.D. Scheme for Electronics and IT. Her research interests include signal processing systems, speech processing, voice conversion, speech synthesis, and human–machine interaction. She qualified in GATE-2011 and 2015, organized by IIT Madras and IIT Kanpur. She received the GATE Scholarship from MHRD, India, during the M.Tech. degree, from 2012 to 2014.

DEEPA GUPTA was born in 1977. She received the Ph.D. degree in natural language processing (example-based machine translation) from the Department of Mathematics and Computer Application, IIT Delhi, in 2005. She has worked as a Postdoctoral Researcher with FBKIRST (Center for Scientific and Technological Research), Trento, Italy. She is currently an Associate Professor with the Department of Computer Science Engineering, Amrita School of Engineering, Bengaluru, India. She has been guiding Ph.D. and graduate students since 2009. She has given invited talks on machine learning and natural language processing in government funded workshops. She has completed two government funded projects related to text plagiarism detection and speech recognition system for Kannada Language in last five years. She is also involved in consultancy projects with industry. Her research interests include sentiment analysis, clinical data mining, speech processing, and other areas in natural language processing. Her research work is published in journals like *Information Processing and Management*, *Expert Systems With Applications*, IEEE ACCESS, and *International Journal of Speech Technology*.

MOHAMMED ZAKARIAH received the B.Sc. degree in computer science and engineering from Visvesvaraya Technological University, India, in 2005, and the master's degree in computer engineering from Jawaharlal Nehru Technological University, India, in 2007. He is currently a Researcher with the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His research interests include bioinformatics, digital audio forensics, speech processing, cloud computing, multimedia, healthcare, and social media. He has published more than 20 articles in various reputed journals.

YOUSEF AJAMI ALOTAIBI (Senior Member, IEEE) received the B.Sc. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees in computer engineering from the Florida Institute of Technology, FL, USA, in 1994 and 1997, respectively. From 1988 to 1992 and 1998 to 1999, he was with Al-ELM Research and Development Corporation, Riyadh, as a Research Engineer. From 1999 to 2008, he was an Assistant Professor with the College of Computer and Information Sciences, King Saud University, where he was an Associate Professor from 2008 to 2012. Since 2012, he has been a Professor with King Saud University. His research interests include digital speech processing, specifically speech recognition and Arabic language and speech processing.

KAUSTAV BHOWMICK (Member, IEEE) received the B.Tech. degree from the West Bengal University of Technology, Kolkata, India, in 2005, and the master's and Ph.D. degrees from the Department of EEE, University of Nottingham, U.K. Primarily as a researcher in photonics, he shares interest in certain medical related fields, with a zeal for the benefit of poor. He was awarded with project from the Ministry of Electronics and IT, Government of India, for working for the benefit of deaf and dumb, as a current venture. He has some experience at NTU Singapore, and has worked at the National Institute of Technology, Sikkim (an Institute of National Importance) and Amrita Vishwa Vidyapeetham Deemed University, as an Assistant Professor. He is currently an Associate Professor with PES University, Bengaluru. His research interests include electromagnetic analyses, photonic and quantum devices, and signal processing systems. He was a recipient of the Nottingham Joint Chevening Scholarship during his master's degree and the Nottingham International Office Research Scholarship during his Ph.D. degree.