Factors Influencing Vocal Pitch in Articulatory Speech Synthesis: A Study Using PRAAT

Sivaramakrishnan Meenakshisundaram
SASTRA University, Thanjavur, India & Interdisciplinary Centre for Computer Music Research, Plymouth University, Plymouth, United Kingdom
sivaramakrishnan.meenakshisundaram@students.plymouth.ac.uk

Eduardo R. Miranda
Interdisciplinary Centre for Computer Music Research, Plymouth University, Plymouth, United Kingdom
eduardo.miranda@plymouth.ac.uk

Irene Kaimi
School of Computing, Electronics and Mathematics, Plymouth University, Plymouth, United Kingdom
irene.kaimi@plymouth.ac.uk

ABSTRACT

An extensive study on the parameters influencing the pitch of a standard speaker in articulatory speech synthesis is presented. The speech synthesiser used is the articulatory synthesiser in PRAAT. Categorically, the repercussion of two parameters: Lungs and Cricothyroid on the average pitch of the synthesised sounds is studied. Statistical analysis of synthesis data proclaim the extent to which each of the variables transform the tonality of the speech signals.

1. INTRODUCTION

A speech synthesiser, outlined in terms of articulatory parameters is a model that incorporates the human vocal tract system. The production of speech sounds using this model is known as articulatory Synthesis. Exploring robust synthesis techniques that can substitute concatenative synthesis has recently become an established field of study. This is due to the limitations that concatenative synthesis imposes on modifying the expressivity of sounds in real time. Articulatory synthesis has been contemplated to have the greatest potential out of the contemporary synthesis techniques [1,2]. Perception of anatomy and physiology of human vocal system is statutory to work with articulatory synthesis. The human vocal tract system can be thought of as a resonating acoustic structure with temporal properties [3]. The entire vocal apparatus is treated as an air filled cavity with walls that may be analogised to adjustable mass-spring systems [4].

The state of the articulatory synthesiser at a particular juncture can be represented by the position of the organs of speech. The time-varying variables of the model simulate a quasistatic speech event that may be represented as a sequence of stationary responses of the model, each corresponding to a particular configuration of the articulatory parameters. It is imperative that the elastic properties of the walls are also considered while defining the vocal tract function.

Articulatory parameters have advantage therein they describe the system that produces the sound instead of the results of that method [1]. However, the irregular form of the vocal tract and temporal properties of the system increase the complexity of modelling.

This project employs the articulatory synthesiser in PRAAT, developed by Paul Boersma and David Weenink. PRAAT is a sophisticated platform for analyzing, synthesizing and manipulating speech [5]. PRAAT comes with its own scripting language and an ideal user interface for analysis and production of speech signals.

The articulatory synthesiser in PRAAT offers 29 degrees of freedom, each typifying an organ/articulatory parameter of the vocalisation system. These parameters of the synthesiser are excited by passing numerical values as input. The physical model in PRAAT provides appreciable realism and naturalness of the sound synthesised, increasing the prospects of implementation of text-speech systems based on articulatory synthesis in the near future.

This paper presents a comprehensive study on factors influencing vocal pitch in articulatory synthesis using PRAAT. The two main parameters that alter the pitch of the human voice are air flow from the Lungs and the Vocal Fold Tension[4]. The cynosure is on the Cricothyroid parameter of the model that is related to the Vocal Fold Tension. This variable is extrapolated beyond the nominal range to observe for changes in the pitch of the sounds synthesised. The results acquired are motivational for further explorations in this discipline.

2. METHODOLOGY

PRAAT features a sophisticated synthesiser that is capable of producing realistic vocal sounds of great interest to composers and artists [6]. For this exploratory measure, the physical model is constrained to 6 parameters to reduce complexity. The model is configured to a standard speaker with two tubes in the glottis. In PRAAT parlance, the Artword Object that encloses all the muscle components. These components can be excited either by directly modifying the Artword or by using the scripting tool in PRAAT. The Artword can be created from the main menu or by using the PRAAT script. The entire set of operations is done on an "A" vowel sound synthesised

using the model. In theory, the parameters of the articulatory synthesiser can vary from -1.0 to +1.0, but in most of the cases we employ 0.0 as the starting point [6].

2.1 Excitation of Parameters

Vocal fold oscillation eventuates in an event of speech. This process can be explained with the Myoelastic-Aerodynamic theory. According to the theory, Bernoulli forces create a closed airspace below the glottis by sucking the vocal folds together. Once the subglottal pressure is high enough, the folds are blown outward causing phonation.

Under conditions of zero vocal fold collision and idealized flow in the glottis the intraglottal pressure can be written as [7],

$$P_g = \left(1 - \frac{a_2}{a_1}\right)(P_s - P_i) + P_i \quad (1)$$

where a_1 and a_2 are the cross-sectional areas at the entry and exit points of the glottis respectively, P_s is the subglottal pressure and P_i is the input pressure to the vocal tract. The term $(P_s - P_i)$ represents the transglottal pressure. Eqn. 1 clearly illustrates the process of vocal fold oscillation resulting in phonation.

This preceding section elucidates the configuration of the muscle parameters enfolded in the Artword, with numerical values that can produce a significant utterance for analysis. This is a time domain approach of simulating the dynamic properties of a human vocalisation system.

The Lungs parameter in the model produces the necessary air pressure to cause phonation. This parameter can be set to attain values between -0.5 and +1.5: the value -0.5 represents the maximum volume of air exhaled by the speaker and the value +1.5 represents the maximum volume of air inhaled [2].

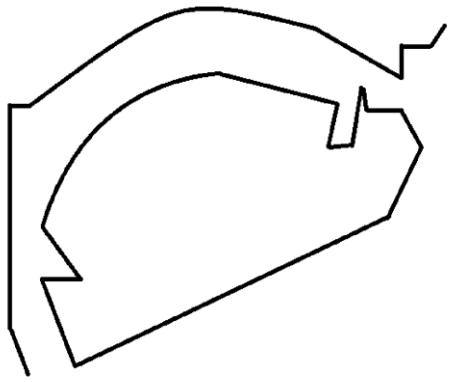

Figure 1. Expiration in PRAAT's articulatory synthesiser

Figure 1 shows the Exhale operation of the articulatory synthesiser in PRAAT. The Exhale operation can be implemented by reducing the equilibrium width of the Lungs.

Adduction of vocal folds is due to the Interarytenoid muscle that connects the two paired, pyramidal Arytenoid cartilages [8]. It consists of two components: the Transverse Arytenoid and the Oblique Arytenoid. These muscles run horizontally across each other forming a shape of letter 'X". This parameter of the synthesiser is set to 0.5 throughout the utterance to produce a normal voicing.

Masseter is a thick rectangular muscle that carries out mastication [9]. This criterion of the model enables opening and closure of the jaw. For an "A" vowel to be synthesised, the jaw should be open. Hence the variable is configured to have a value of -0.4 throughout the utterance. Similarly, a value of 0.5 can be set for complete jaw closure. The Hyoglossus, one of the extrinsic muscles of the tongue is quadrangular in shape in the lower two-thirds and radiates into fan shaped structure in upper one-third [10]. This muscle is responsible for depression and retraction of human tongue and is a cardinal muscle in singing. In order to produce an "A" sound, this muscle variable is excited with a value of +0.4 throughout the utterance.

Cricothyroid is a muscle that influences the pitch of the sound produced in a human vocal tract by altering the tension and the length of the vocal folds [4,6]. This muscle stretches forwards and backwards to modulate the pitch. Akin to tightened guitar string, the stretched Cricothyroid muscle produces a high-frequency sound. The Cricothyroid parameter of the synthesiser is of great interest as recent experiments have shown the muscle's direct relation to the fundamental frequency of the human voice [11]. This parameter is excited with values between 0 and 4 to observe for changes in the vocal pitch of the synthesiser.

PRAAT has a great advantage in producing sounds such as nasals to a substantial extent of realism. The LevatorPalatini muscle parameter of the synthesiser is responsible for opening and closure of the Velopharyngeal port. Oral phonemes require the port to be closed and the nasal phonemes are produced by nasal resonance, which is accomplished by opening the Velopharyngeal port and allowing the air to pass through the nasal cavity. This operation in PRAAT can be achieved by exciting the LevatorPalatini variable. In this project, as an "A" vowel is synthesised, a value of +1.0 is assigned to the parameter for enabling complete closure of the Velopharyngeal port. A value of 0.0 enables opening of the port.

2.2 Synthesis

PRAAT is highly flexible that it is adequate to just specify the discrete settings of the parameters. It automatically introduces the values in between the discrete levels. A sustained phonation is achieved by setting the duration of the utterance to 1.5s. This gives a wider window to analyse the behaviour of the pitch throughout the utterance. The time-varying variables of the model are excited at discrete time levels that are within the length of the phonation.

Every variable of the Artword can be excited through PRAAT script. For example the statement: Set target: 0.03, -0.1, "Lungs", initializes the Lungs parameter with a value of -0.1 at the time instant 0.03. All other parameters are excited in similar fashion.

Once all the variables are initialized, the Artword along with the Speaker properties is synthesised to a sound. PRAAT offers 9 different options (Sampling frequency, Oversampling factor, Width 1, Width 2, Width 3, Pressure 1, Pressure 2, Pressure 3, Velocity 1, Velocity 2, Velocity 3) to synthesise the sound. The sound synthe-

sised is directly related to the muscles that are configured in the model.

All results presented in this paper are based on the "Average Pitch" of the sound, as the pitch tends to vary along the utterance. Not all parameter configurations result in a normal voicing. Some of the configurations tend to produce voiceless sounds that don't have a definite pitch. There may also be breaks in the sounds synthesised. It is difficult to say if these breaks are related to the realistic phenomena of the model [2]. Another noteworthy characteristic of the model is that the sounds relating to a particular configuration, when re-synthesised after a certain interval of time result in different tonal and spectral properties, manifesting the imperfect nature of the physical model in PRAAT. Aforementioned qualities attribute to the unexploited nature of articulatory speech synthesis in robust applications.

3. RESULTS

In order to get an insight into the functioning of the model, a very large set of sound samples is required. Each of these sound samples is related to a particular configuration of the Artword parameters. Four of the six parameters are kept constant throughout the experiment. For this experimental study, the model is simulated with different combinations of air flow and vocal fold tensions along with the other four static parameters. This results in 450 different speech sounds of same phonation length with different tonal and spectral properties. These sounds are statistically analysed to find the effect of these variables on the average pitch of the speaker to which the synthesiser is configured.

3.1 Histogram

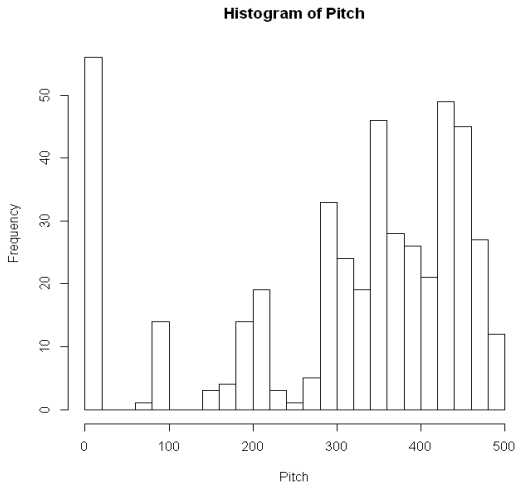

Figure 2. Histogram of Vocal Pitch (Average) in Hz data corresponding to different configuration of Lungs and Cricothyroid

The Histogram of the variable pitch is first plotted. As Figure 2 suggests, the distribution of pitch is not symmetric (due to the large number of zero's). This suggests that normality of the variable cannot be assumed.

Then a formal statistical test for normality (a Shapiro-Wilk test) is done. The p-value of the test is 0.000, hence the null hypothesis of normality is rejected. As a result, parametric tests cannot be performed hence, non-parametric methods to assess association of pitch with the available covariates are employed.

3.2 Effect of Exhale Levels on Resultant Pitch

In PRAAT, the Lungs parameter has a greater impact on the amplitude of the sounds synthesised. In addition to this, it also influences the pitch of the synthesiser on a smaller scale. This effect should be taken into consideration as they tend to stimulate the Cricothyroid. From Figure 3, it can be seen that the distribution of the pitch is similar for all six levels of Exhale. The median values of the pitch in the six groups (shown as the solid lines inside the boxplots) are approximately at the same value. This suggests that there is minimal Exhale effect on pitch. Then a Kruskal-Wallis test is also performed to statistically assess the Exhale effect on pitch. This gives a p-value=0.798, hence the null hypothesis of the test, that there is no much difference in the median values of the average pitch for the six levels of Exhale, cannot be rejected, which suggests that there is no statistical evidence of an Exhale effect.

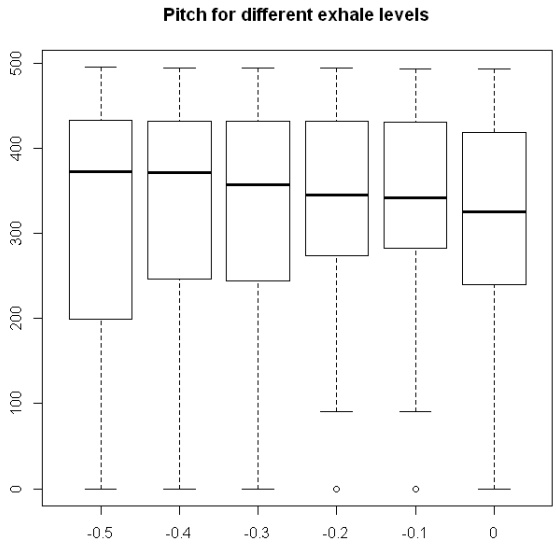

Figure 3. Vocal Pitch (Average) in Hz plot for six levels of Exhale configuration of the Lungs parameter of the synthesiser

3.3 Effect of Cricothyroid Levels on Resultant Pitch

Figure 4 indicates that the distribution of the pitch differs for the six levels of Cricothyroid. For example, at level 0, pitch takes a wide range of values, whereas, at level 4 the range of values of pitch is very limited. The median values of pitch in the six groups (shown as the solid lines inside the boxplots) appear to be very different. For example, at level 0 the median is close to 0, whereas at level 6 the median Pitch value is around 450. This suggests

that there may be a Cricothyroid effect on pitch. Then a Kruskal-Wallis test is done to statistically assess the Cricothyroid effect on pitch. This gives a p-value=0.000, hence there is overwhelming evidence to reject the null hypothesis of the test, that there is no difference in the median values of pitch for the five levels of Cricothyroid. This suggests that there is a strong statistical evidence of a Cricothyroid effect.

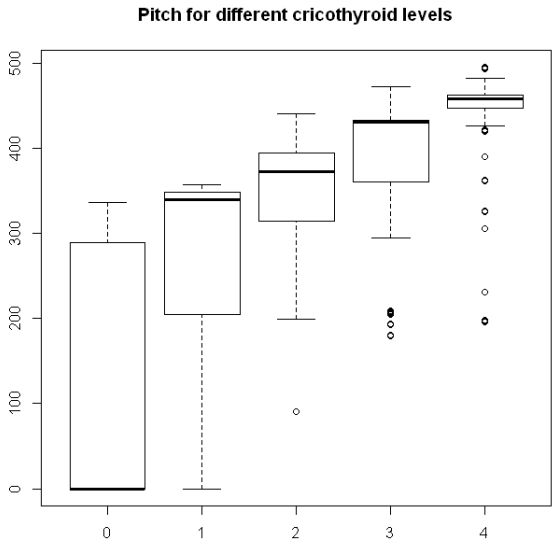

Figure 4. Vocal Pitch (Average) in Hz plot for five levels of Cricothyroid parameter of the synthesiser

3.4 Effect of Inhale Levels on Resultant Pitch

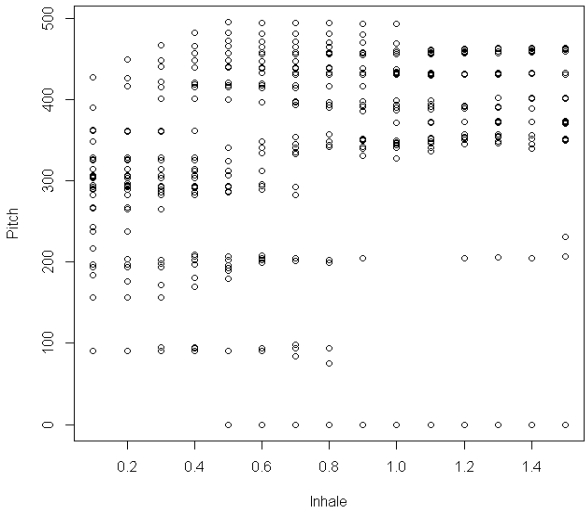

Figure 5. Female Vocal Pitch (Average) in Hz plot for seven levels of Inhale configuration of Lungs parameter of the synthesiser

Figure 5 does not reveal any obvious relationship between pitch and Inhale. This suggests that there is probably no Inhale effect on pitch. A Kendall tau rank correlation coefficient test is performed to statistically assess the Inhale effect on pitch. This gives a low p-value<0.001, hence there is an evidence to reject the null hypothesis of the test, that the correlation between pitch and Inhale is equal to 0. However, the size of the correlation coefficient is 0.19 (very low, compared to 1 which would indicate perfect linear relationship), which suggests that there is no association between pitch and Inhale.

3.5 Non-Parametric Model

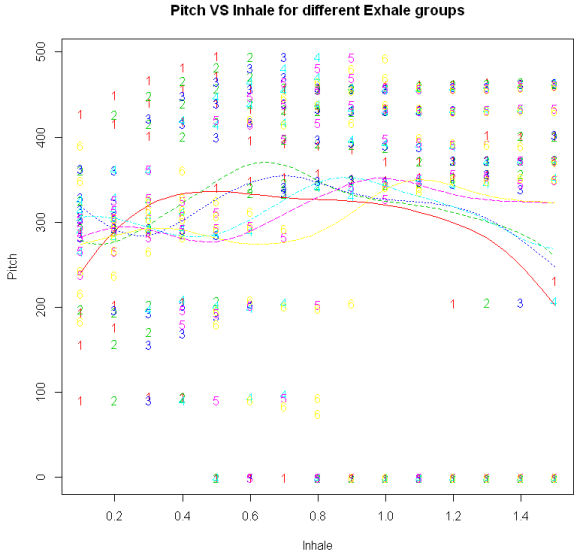

Figure 6. Vocal Pitch (Average) in Hz Vs Inhale plot for different Exhale groups

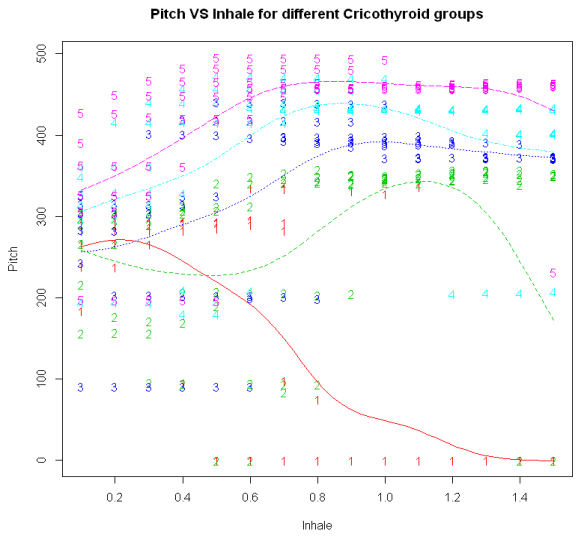

Figure 7. Vocal Pitch (Average) in Hz Vs Inhale plot for different Cricothyroid groups

Finally, the non-parametric equivalent of an ANCOVA model is fitted to the data (in which all the available covariates: Inhale, Exhale and Cricothyroid are included) and their significance and how they affect the response pitch is assessed.

As shown from Figures 6 and 7 , there is no difference between different exhale groups, whereas different Cricothyroid groups differ in how pitch and inhale are associated. The results confirm that there is no association be-

tween Inhale and pitch, and that the pitch medians is the same for all six levels of exhale. However, there is a difference in the median pitch of different Cricothyroid levels. A multiple comparisons provides evidence of statistical differences between all pairs of Cricothyroid levels (0-1,0-2,0-3,0-4,…,4-1,4-2,4-3).

4. CONCLUSIONS AND FUTURE WORK

A detailed analysis of the repercussion of articulatory parameters on female vocal pitch has led to the following conclusions.

- Not all parameter configurations result in a normal voicing. This evinces the realistic phenomena of the articulatory model in PRAAT.
- Inhale pressure levels of the Lungs do not have any effect on the intrinsic vocal pitch of the sounds synthesised using the model.
- The resultant pitch medians of different Exhale levels are approximately at the same range. Thus no statistically significant effect of the Exhale pressure levels on the tonality of the speech sounds synthesised can be identified.
- Divergence in pitch medians for different levels of Cricothyroid, reinforces the fact that vocal cord tension is directly related to the intrinsic vocal pitch of the speech signals. Statistical studies also arrive to a conclusion that the Cricothyroid has a solid role in altering the pitch of the synthesiser.

Future work will go towards exploring other articulatory parameters of the model that transform the tonality of the speech sounds. To initiate, the Thyroarytenoid and the Vocalis parameters of the model will be studied extensively along with the other six parameters used in this experiment. Further explorations will lead the way to discover parameter configurations for different vowels and consonants, transition between individual sounds using the model.

5. REFERENCES

[1] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," 2002.

[2] J. Drayton and E. Miranda, "Towards an Evolutionary Computational Approach to Articulatory Vocal Synthesis with PRAAT," in *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, ed: Springer, 2015, pp. 62-70.

[3] C. Wu and Y.-F. Hsieh, *Articulatory speech synthesizer*: University of Florida, 1996.

[4] P. Boersma, "Functional phonology," ed: The Hague: Holland Academic Graphics, 1998, pp. 31-63, pp. 113-140.

[5] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT," *Glot International,* vol. 5, 2001, pp. 341-347.

[6] E. Miranda, *Computer sound design: synthesis techniques and programming*: Taylor & Francis, 2012, pp. 137-152.

[7] M. A. Redford, *The handbook of speech production*: John Wiley & Sons, 2015, pp. 34-58.

[8] C. A. Rosen and B. Simpson, *Operative techniques in laryngology*: Springer Science & Business Media, 2008, pp. 3-8.

[9] T. Van Eijden, "Jaw muscle activity in relation to the direction and point of application of bite force," *Journal of dental research,* vol. 69, 1990, pp. 901-905.

[10] S. Abd-El-Malek, "Observations on the morphology of the human tongue," *Journal of Anatomy*, vol. 73, 1939, pp. 201-210.

[11] D. Erickson, T. Baer, and K. S. Harris, "The role of the strap muscles in pitch lowering," *Vocal fold physiology: Contemporary research and clinical issue*,1983, pp. 279-285.