Interspeech 2018
2-6 September 2018, Hyderabad

Implementation of Respiration in Articulatory Synthesis Using a Pressure-Volume Lung Model

Keisuke Tanihara[1], *Shogo Yonekura*[1], *Yasuo Kuniyoshi*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan

tanihara@isi.imi.i.u-tokyo.ac.jp, yonekura@isi.imi.i.u-tokyo.ac.jp,
kuniyosh@isi.imi.i.u-tokyo.ac.jp

Abstract

In previous studies of the 1D vocal tract model of articulatory synthesis, subglottal pressure is typically regarded as constant, ignoring its dynamics. However, human vocalization is initially generated by glottal airflow via subglottal pressure change. This change is caused by the expansion and contraction of the lungs. In the current study, we propose a new pressure-volume model that relates pressure changes to volume changes of the human lung. Using this model, the behavior of the human lung can be integrated with articulatory synthesis. This model produces positive and negative subglottal pressure corresponding to expiration and inspiration respectively. In addition, breathing could be implemented in the proposed model. This implementation would expand the possibilities for articulatory synthesis.

Index Terms: speech synthesis, articulatory synthesis, lung, time domain simulation

1. Introduction

Articulatory synthesis is one of speech synthesis approaches. This approach simulates human speech production processes using a mathematical model of the speech organs. Conventional synthesis approaches, such as concatenative synthesis [1] reconstruct real human voice and have the advantage of naturalness. However, they ignore the body conditions underlying human voice production, and are less related to actual vocalization movements. Compared with such approaches, articulatory synthesis that uses mathematical models of the speech organs can be more easily related to actual vocalization movements because the models directly correspond to each organ.

The 1D vocal tract model has been used in many previous studies of articulatory synthesis [2, 3, 4]. This model regards the vocal tract as connected cylinders, and assumes plane wave propagation. Conventional articulatory synthesis systems using this model consider the vocal folds, vocal tract, and nasal tract. However, a few studies have considered the lungs. Some studies considering the lungs [4] have assumed a static lung shape, and other studies [5, 6] have assumed varying lung shape but considered only expiration and ignored inspiration into lung via glottis.

In actual vocalization, expansion and contraction of the lungs generate subglottal pressure change and glottal airflow. This airflow then produces vocal folds vibration and vocal tract resonance. In addition, subglottal pressure and the fundamental frequency of voice are thought to have a linear relationship [7]. Thus, the effects of subglottal pressure on speech sound are considered to be important. Nevertheless, most of previous studies of articulatory synthesis systems consider subglottal pressure as a constant or only expiration, and ignore changes in the shape of the lungs or inspiration.

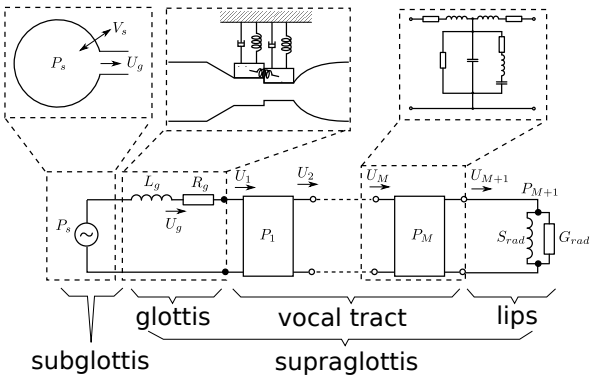

Figure 1: *System components*

To solve this problem, the current study proposes an simple approximate lung model, which generates subglottal pressure by quasi-static and adiabatic volume changes of the lungs. Positive and negative subglottal pressure, which corresponds to expiration and inspiration, were created with the proposed lung model. Breathing in vocalization, which has not been realized in previous articulatory synthesis models, can be implemented by this subglottal pressure change in the proposed model.

2. Model

Figure 1 shows the components of the proposed system. The proposed system consists of a supraglottal articulatory synthesis model and a subglottal pressure-volume model. The supraglottal model includes the vocal folds and vocal tract, while the subglottal model is a pressure-volume lung model.

2.1. Supraglottal Articulatory Synthesis Model

The 1D vocal tract model and plane wave propagation were adopted as the articulatory synthesis system in the supraglottal model, which were the same as the model used by Ho et al.[4] The two mass model proposed by Ishizaka and Flanagan [8] was used as a glottal model. We calculated self-oscillation of the vocal folds generated by the difference between supraglottal pressure and subglottal pressure. Sound propagation was calculated using the transmission line circuit model [2] which is equivalent to the 1D vocal tract model.

Although more sophisticated models exist in glottis [9, 10] and vocal tract [11, 12], simple models [2, 4, 8] were selected for the this study. The simple models are adequate for realizaion expiration and inspiration, and the complicated models are unnecessary for this case.

10.21437/Interspeech.2018-1080

2.2. Subglottal Pressure-Volume Model

The simple lung model proposed in this study is described by two parameters: subglottal pressure P_s and subglottal volume V_s. The relationship between pressure change and volume change is expressed in (1) when the time step in the simulation was short enough to consider quasi-static and adiabatic conditions of volume change. η is the heat capacity ratio given by $\eta = 1.4$, regarding air as a diatomic molecule.

$$PV^{\eta} = P'V'^{\eta} \tag{1}$$

When subglottal pressure and volume are described as P_s, V_s, they become $P'_s, V_s + \Delta V$ after time step Δt. Substituting these parameters for (1) then yields to (2).

$$\begin{aligned} P_s V_s^{\eta} &= P'_s (V_s + \Delta V)^{\eta} \\ P'_s &= P_s \left(\frac{V_s}{V_s + \Delta V} \right)^{\eta} \end{aligned} \tag{2}$$

We considered that subglottal volume displacement ΔV after a time step Δt would be sufficiently smaller than subglottal volume V_s ($|V_s| \gg \Delta V$). Second or higher order terms of Taylor expansion around $\frac{\Delta V}{V_s} = 0$ can then be ignored, and the following equation can be established.

$$\begin{aligned} \left(\frac{V_s}{V_s + \Delta V} \right)^{\eta} &= \left(1 + \frac{\Delta V}{V_s} \right)^{-\eta} \\ &\simeq 1 - \eta \frac{\Delta V}{V_s} \end{aligned} \tag{3}$$

Substituting this equation for (2) becomes the first-order approximation equation (4).

$$P'_s \simeq P_s \left(1 - \eta \frac{\Delta V}{V_s} \right) \tag{4}$$

The full dynamics of subglottal pressure and volume is given by (2) and its first-order approximation is given by (4).

The 1D vocal tract model uses the superposition principle for plane waves, and uses the difference value from atmospheric pressure as the pressure value. However, the pressure-volume relationship in (2) is not linear, making the principle invalid. Thus, a pressure value that includes atmospheric pressure instead of a difference value is necessary in the proposed lung model. To transfer P_s from a subglottal model to a supraglottal model requires the subtraction of atmospheric pressure.

The subglottal volume is controlled as follows using the target subglottal pressure P'_s. Using (2) and its first-order approximation (4), the manipulation of subglottal volume displacement ΔV can be calculated as (5) and (6) using target subglottal pressure P'_s.

$$\Delta V = V_s \left(\frac{P'_s}{P_s} \right)^{\eta} - V_s \tag{5}$$

$$\simeq \frac{V_s}{\eta} - \frac{V_s P'_s}{\eta P_s} \tag{6}$$

The volume change ΔV can be separated subglottal volume change ΔV_s due to shape change, and outflow volume at the glottis $U_o \Delta t$ connected to the supraglottal model in simulation ($\Delta V = \Delta V_s + U_o \Delta t$). The manipulated volume value ΔV_s for target subglottal pressure P'_s can then be calculated using the following equations (7) and (8).

$$\Delta V_s = V_s \left(\frac{P'_s}{P_s} \right)^{\eta} - V_s - U_o \Delta t \tag{7}$$

$$\simeq \frac{V_s}{\eta} - \frac{V_s P'_s}{\eta P_s} - U_o \Delta t \tag{8}$$

2.3. Calculation of the System

Using the above supraglottal and subglottal system, the output sound can be calculated using the following iteration. The manipulation value of subglottal volume V_s in the expiration phase is calculated based on the target subglottal pressure using (7) and (8). And that in the inspiration phase is also calculated based on the end-inspiratory position (EIP), which is the sum of the functional residual capacity and tidal volume. EIP is set as the target volume and compared to the maximum of volume velocity based on $\mathrm{FEV}_1\%$. $\mathrm{FEV}_1\%$ means exhaled air volume ratio of forced expiration in the first 1 second.

1. Calculates subglottal volume changes
   (a) subglottal volume changes by shape transform into $V'_s = V_s + \Delta V_s$
   (b) Volume $U_o \Delta t$ flows out of glottis
2. subglottal pressure changes into $P'_s = P_s \left(\frac{V_s}{V_s + \Delta V_s + U_o \Delta t} \right)^{\eta}$
3. The lung is connected to the equivalent supraglottal circuit as a pressure source P'_s
   (a) Calculates vibration displacement of two mass model
   (b) Simulates volume velocity and pressure in vocal tract equivalent circuit
   (c) Gets output sound pressure P_{out} and glottal volume velocity U_g
4. Goes back to 1

In this iteration, the system is operated by appropriate changes of subglottal volume V_s. For simplicity, a direct connection between the subglottis and glottis is assumed, and airflow from the subglottis U_o is equal to glottal airflow U_g.

2.4. Numerical Calculation Method

The synthesis system is calculated in the time domain. The classical Runge-Kutta method is used to simulate the two mass model. Slight displacement of glottal volume velocity dU_g is calculated by a trapezoidal rule, which was also used in the vocal tract in Maeda's system [2]. In addition, attenuation coefficient $\alpha (< 1.0)$ is introduced to this accumulation term. This coefficient prevents noise generation in respiration experiments. The relationship of U_g, dU_g is shown in the next equation (9).

$$\begin{cases} dU_g[n] = 2 \frac{U_g[n]}{\Delta t} - Q_{U_g}[n-1] \\ Q_{U_g}[n] = 4 \frac{U_g[n]}{\Delta t} - \alpha Q_{U_g}[n-1] \end{cases} \tag{9}$$

2.5. Simulation Parameters

Parameters used in the simulation are described here.

The area function of the vocal tract reflects the magnetic resonance imaging measurement result by Story et al.[13] and the function of a vowel /a/ was adopted. The glottal parameters in the two mass model and the air parameters were the same as those in the system used by Ho et al.[4] The attenuation coefficient α in the proposed model was set to 0.999.

The initial subglottal pressure was set as the atmospheric pressure (1.013×10^6 dyn/cm^2). The initial subglottal volume was also set as EIP, and its approximation value was 3.0×10^3 cm^3. The maximum subglottal volume velocity was defined as FEV$_1$%. Berglund FEV$_1$%[14] was calculated as 4.22×10^3 under conditions in which height was 170.9 cm (the average height of a 20-year-old Japanese male). The calculated volume velocity ΔV_s was based on (7) and (8) was transformed into volume velocity in atmospheric pressure by (2) or (4), then compared to FEV$_1$%.

3. Experiments

In the following section we describe three experiments, including presentation of the results and discussion of the findings. The operating frequency of the system was set to 5.0×10^5 Hz. For calculations, we examined the subglottal pressure-volume relationship based on quasi-static and adiabatic conditions using equation (2) and its first-order approximate equation (4), and compared the results.

3.1. Subglottal volume control

The first experiment was conducted to examine subglottal volume control compared with constant subglottal pressure. Target subglottal pressure was set as 8 cmH$_2$O(=7.843×10^3 dyn/cm^2). This pressure value has been used in many previous studies of articulatory synthesis systems, which have typically used constant subglottal pressure. The results of the proposed model were compared with the results of a conventional system that replaces the subglottal part of the proposed system by such constant subglottal pressure value.

3.2. Different target subglottal pressure

In the second experiment, the proposed lung model operated with different target subglottal pressure values and controled its volume. These values were set as 4, 6, 8, 10, 12 cmH$_2$O. By controlling subglottal volume, we changed subglottal pressure to trace target pressure values. We compared differences in the fundamental frequency of these output sound. When the effects of the proposed lung model on output were confirmed, we assumed that the use of subglottal parameters in addition to supraglottal parameters would make articulatory synthesis more natural and complex.

3.3. Implementation of inspiration & expiration

In the third experiment, we confirmed whether aspiration was realized by extension and contraction of the proposed lung model. Breathing in vocalization was assumed, and 1.5 s vocalization consisted of 0.5 s expiration, followed by 0.5 s inspiration, and 0.5 s expiration. The target subglottal pressure in expiration was set as 8 cmH$_2$O. In inspiration, displacement and velocity of the two mass model was fixed as $x_i = 0, v_i = 0$. Thus, the glottis was slightly opened. Because previous articulatory synthesis systems have not realized inspiration action, implementation of the action would expand the application range of the system.

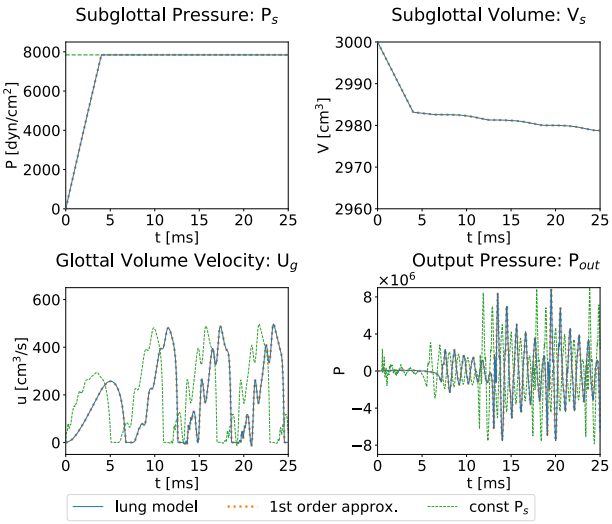

Figure 2: *Subglottal pressure, volume, glottal volume velocity, and output pressure generated by pressure-volume lung model and constant subglottal pressure*

4. Results and Discussion

4.1. Subglottal volume control

Figure 2 shows the results as the time varying output of subglottal pressure P_s, subglottal volume V_s, glottal volume velocity U_g, output sound pressure P_{out}. In the proposed system, the results of the pressure-volume relationship using equation (2) and its first-order approximate equation (4) closely corresponded. The blue solid line shows the results of the proposed pressure-volume model, the orange dotted line shows the results of the approximate equation, and the green dashed line shows the results of constant pressure. Both the blue and orange lines in P_s showed a maintained increase toward target subglottal pressure by a decrease in V_s. Consequently, similar output U_g, P_{out} to output of constant pressure suggests that the synthesis system using the proposed lung model was able to generate self-oscillation of the vocal folds, and resonance in the vocal tract. In addition, the gradual increase in the target value in P_s resulted in a smoother (and maybe more natural) start of the waveforms of U_g, P_{out} compared with those generated with constant lung pressure.

4.2. Different target subglottal pressure

Similarly to the first experiment, each subglottal pressure pattern in the second experiment tracked the target pressure pattern. The fundamental frequency of the results was calculated using the autocorrelation method, and the findings are shown in Figure 3. The blue dots show the pitches calculated from the pressure-volume equation (2), and the orange "x" symbols show the pitches calculated from its first-order approximate equation (4). The frequency values of each equation were the same. The green dashed line shows the approximate line calculated using the least squares method. As the target pressure increased, the fundamental frequency also increased. The relationship between these values appeared to exhibit a linear pattern, and the

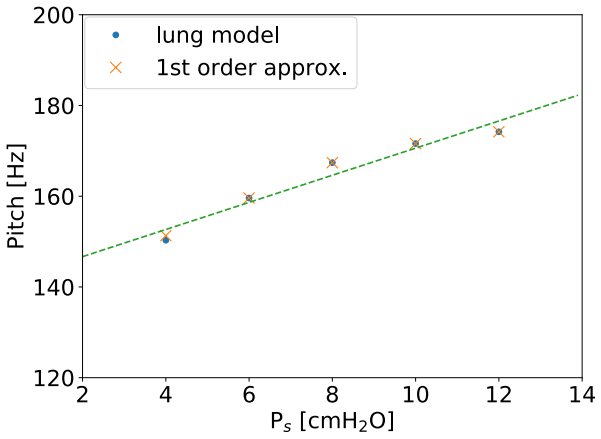

Figure 3: *Fundamental frequency produced by different subglottal pressure*

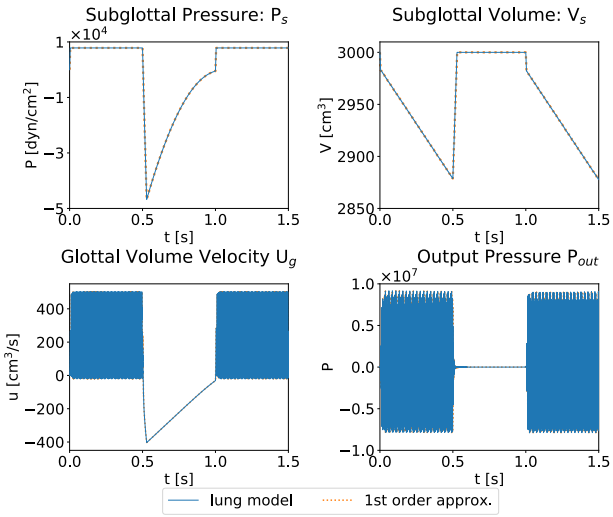

Figure 4: *Subglottal pressure, volume, glottal volume velocity, and output pressure among expiration and inspiration with attenuation coefficient*

proportion corresponded to the estimation by Titze[7].

4.3. Implementation of inspiration & expiration

Examining the generated sound confirmed that the vowel /ɑ/ was produced in two expiration phases, while the inspiration phase was silent in both pressure-volume equations (2) and its first-order approximate equation (4). Figure 4 describes the time varying output of subglottal pressure, subglottal volume, glottal volume velocity, and output sound pressure in expiration and inspiration. The blue solid line corresponds to the proposed pressure-volume model, while the orange dotted line corresponds to the approximate equation of the model. Decreased subglottal volume in expiration phase recovered in the inspiration phase, and negative glottal volume velocity led to recovery of subglottal pressure. These results suggest that the implementation of expiration and inspiration (i.e., respiration) was realized by the proposed pressure-volume lung model. To implement the aspiration sound involved in the frictional sound in the silent inspiration phase mentioned above, the following two factors should be considered. The first factor is the consideration of frictional sound and the extension of the articulatory model we used, which did not produce this sound. The second factor is the development of a subglottal volume control method in the inspiration phase based on actual measurement, to make U_g, P_s more natural during inspiration.

In addition, attenuation coefficient α was required in this experiment. The unstable vibration in U_g, P_{out} of the second expiration phase was generated without this coefficient (i.e., $\alpha = 1.0$). The output sound in the second expiration phase included noise, and cannot be regarded as a regular vowel. The cause of this noise was likely to be related to the form of the accumulation term. First-order derivation of U_g did not change in the inspiration phase compared with the expiration phase. Because the sum of the accumulation term consisted of the sum of the first-order derivation, the accumulation term became overly large during inspiration. Thus, the term was not able to reflect small fluctuations of U_g in the second expiration, and generated an unstable wave.

In the first and second experiments, the output of pressure-volume equation (2) and its first-order approximate equation (4) were approximately identical. In the third experiment, the behavior of P_s, V_s, U_g, P_{out} was almost identical, and no difference could be heard when listening to the sounds. This suggests the minimal effect of the approximation on the output sound quality in practical usage. The calculation time of present synthesis system was largely occupied by the propagation process in the vocal tract, and relatively little time was taken up by the lung occupancy component. Thus, the results indicated that use of the approximation in the articulatory synthesis system was not necessary. However, future extensions of the proposed lung model, such as consideration of the alveolus and branches, would be expected to increase the amount of calculation involved. Thus, first-order approximation may have the advantage of decreasing calculation with minimal effects on the output sound.

5. Conclusions

In the current study, a new pressure-volume model of the human lung was proposed and integrated into an articulatory synthesis system. Various equations of the pressure-volume relationship in the lung model were considered. The results revealed that the first-order approximation had the minimal effect on the output sound quality in practical usage.

This implementation required attenuation coefficient in accumulation terms. Without attenuation, changes of the first derivation of volume velocity in the inspiration phase were small, meaning that the accumulation term became overly large and was likely to have induced noise.

subglottal pressure in the proposed pressure-volume model was controlled by changing subglottal volume, and had an effect on the fundamental frequency of output sounds. Inspiration in vocalization was realized by consideration of the physical conditions of the lung. Unlike previous synthesis systems, the proposed system implemented aspiration actions, extending the development of articulatory synthesis.

6. Acknowledgements

We thank Benjamin Knight, MSc., from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

7. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, 1996, pp. 373–376.

[2] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech communication*, vol. 1, no. 3, pp. 199–229, 1982.

[3] P. Mokhtari, H. Takemoto, and T. Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches," *Speech Communication*, vol. 50, no. 3, pp. 179–190, 2008.

[4] J. C. Ho, M. Zañartu, and G. R. Wodicka, "An anatomically based, time-domain acoustic model of the subglottal system for speech production," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1531–1547, 2011.

[5] P. Boersma, "Functional phonology: Formalizing the interactions between articulatory and perceptual drives," Ph.D. dissertation, University of Amsterdam, 1998.

[6] ——, "Interaction between glottal and vocal-tract aerodynamics in a comprehensive model of the speech apparatus," in *Proceedings of the International Congress of Phonetic Sciences*, vol. 2, 1995, pp. 430–433.

[7] I. R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 901–906, 1989.

[8] K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords," *Bell system technical journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[9] P. Birkholz, B. J. Kr, and C. Neuschaefer-rube, "Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds," in *First International Workshop on Performative Speech and Singing Synthesis. 2011*, vol. 370, 2011.

[10] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, p. 1249, 1995.

[11] B. Elie and Y. Laprie, "Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink," *Speech Communication*, vol. 82, pp. 85–96, 2016.

[12] P. Birkholz, "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.

[13] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[14] E. D. Baldwin, A. Cournand, and D. W. Richards Jr., "Pulmonary insufficiency; physiological classification, clinical methods of analysis, standard values in normal subjects," *Medicine*, vol. 27, no. 3, pp. 243–278, 1948.