Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Development of speech synthesis simulation system and study of timing between articulation and vocal fold vibration for consonants /p/, /t/ and /k/

Kohichi Ogata (1) and Mamio Hokazono (1)

(1) Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan

PACS: 43.72.Ja, 43.70. Bk

ABSTRACT

This paper describes development of an articulatory speech synthesis system using a transmission line model. A speech synthesis method that simulates speech production process has a potential to produce human-like speech. However, there are many parameters to be set such as vocal tract area functions and timing between articulatory movement and vocal fold vibration. A simulation system that allows us to evaluate the effects of related parameters on synthesized speech provides us with useful information on adequate values for the related parameters. In this paper, therefore, an articulatory speech synthesis system with graphical user interface (GUI) has been developed to overcome the difficulty in setting control parameters. The system is based on the speech synthesizer proposed by Sondhi and Schroeter. In the synthesizer, the two-mass model of the vocal folds proposed by Ishizaka and Flanagan is used to produce the glottal source. This paper focuses on producing of stop consonants such as /p/, /t/ and /k/. The GUI-based simulation system has been developed in Java language in order to investigate relative timing between articulation and vocal fold vibration. First, area functions for /p/, /t/ and /k/ were studied by evaluating transfer functions for acoustic tubes as vocal tract shapes, because the difference of frequency range of energy distribution among these consonants is one of important cues. Secondly, relative timing between the events of articulation and vocal fold vibration was studied by evaluating its effects on the synthesized speech from the point of view of voice onset time (VOT). The simulation results showed adequate ranges between both the events for producing successful consonant sounds /p/, /t/ and /k/. Moreover, these results allow us to utilize the timing information for improving the simulation system having an automatic adjustment function for setting related parameters.

INTRODUCTION

A speech synthesis method that simulates speech production process; generation of sound source, articulation and radiation has a potential to produce human-like speech. Because the speech output is calculated through the physiological model of the speech production, this method has advantages in producing smooth continuous speech resulting from inherent continuity of articulatory parameters. However, controlling articulatory parameters is not easy because of a number of muscles involved in the articulation and the complexity in control strategy.

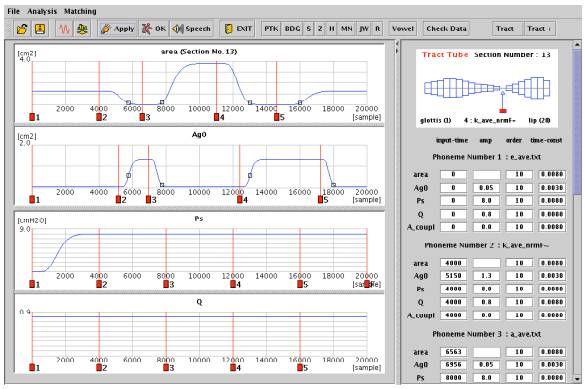
Utilizing the continuity of the vocal tract shape for synthesizing natural continuous speech, the authors have developed a speech synthesis system using a transmission line model [1]. In the system, a vocal tract is modelled as 20 acoustic tubes and the change in the areas of the acoustic tubes as a function of time is described as the time patterns of the step response of cascaded first order systems [2].

A simulation system that allows us to evaluate the effects of related parameters on synthesized speech provides us with useful information on adequate values for the related parameters. In this paper, therefore, an articulatory speech synthesis system with graphical user interface (GUI) has been developed to overcome the difficulty in setting control parameters.

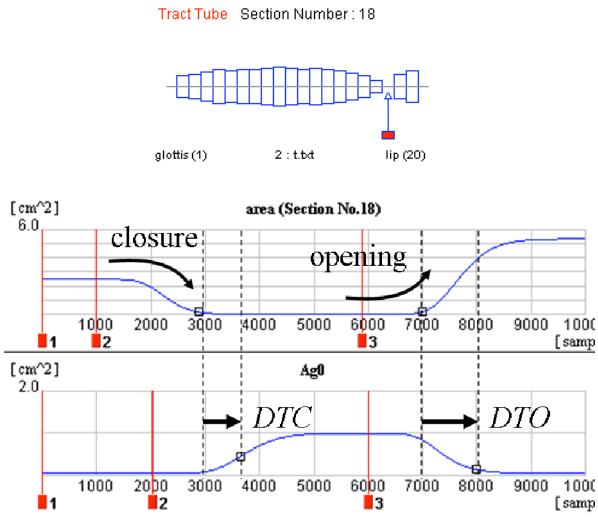
SPEECH SYNTHESIS SYSTEM

This section shows an overview of the speech synthesis system and simulation for relative timing between articulation and vocal fold vibration. The speech synthesizer used in our system is based on the speech production model proposed by Sondhi and Schroeter [3]. In the synthesizer, the two-mass model of the vocal folds proposed by Ishizaka and Flanagan [4] is used to produce the glottal source. The excitation of the glottis is controlled by vocal-fold tension (Q), glottal rest area (A_{g0}), and subglottal pressure (P_s). In our system, the vocal tract is approximately represented by 20 cascaded acoustic tubes and is parameterized by cross-sectional areas of the tubes and the vocal tract length as shown in the upper right of Figure 1(a).

Describing a continuous change in the vocal tract area functions is necessary for the synthesis of continuous speech. In our synthesis system, the continuous change is represented by changing the area of each acoustic tube based on the behavior of the cascaded first-order systems [2]. Because trajectories of articulatory movements were closely approximated by the

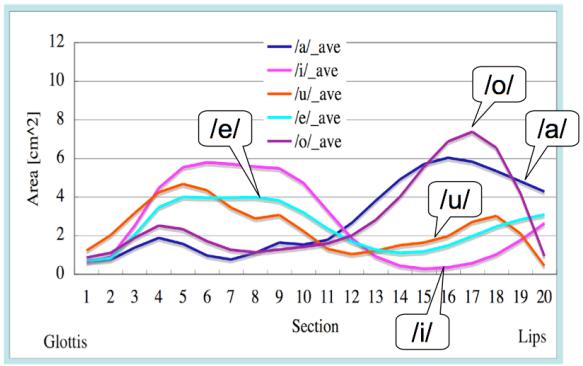


(a)



(b)

Figure 1: (a) Snapshot of a GUI window and (b) definition of the time intervals *DTC* and *DTO*.



Vocal tract length [cm]				
/i/	/e/	/a/	/o/	/u/
16.97	16.36	16.89	17.37	17.81

Figure 2: Vocal tract area functions for five vowels used in the simulation.

response pattern of the cascaded first-order systems in the previous experiments [2], the response pattern is used for describing the continuous change.

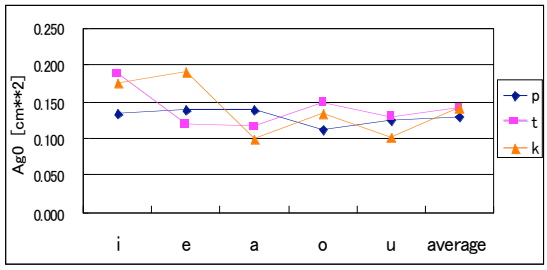


Figure 3: Theshold for A_{g0}

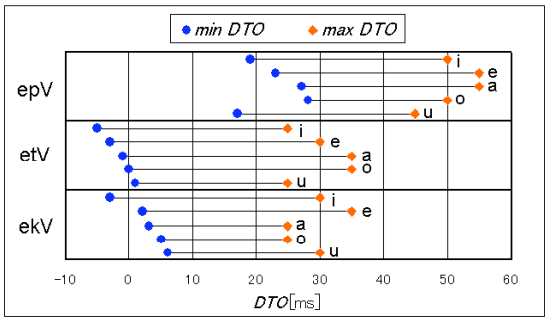


Figure 4: Time interval *DTO*s for phonetic sequence /eCV/(C= p, t, k, V= i, e, a, o, u).

GUI window and definition of time intervals

Figure 1 (a) shows a GUI window for speech synthesis. At the top of the figure, a change in the area of one acoustic tube, which is the 13th section indicated by an arrow in the vocal tract view window, is shown as a function of time. The change in glottal rest area (A_{g0}) that controls voiced or voiceless speech is shown in the second row. The changes in subglottal pressure (P_s) and vocal-fold tension (Q) are also shown in the third and fourth rows, respectively.

The definition of the relative timing between articulation and vocal fold vibration is shown as time intervals *DTC* and *DTO* in Figure 1 (b). Each interval is defined as a time interval divided by small squares on the curves that represent thresholds for the area and the glottal rest area (A_{g0}). The detail of these thresholds will be described later. Long vertical lines with numbers indicate the time instants at which step inputs as hypothetical motor commands to the cascaded fist-order systems occur. Because these instants can be changed with a computer mouse, various time intervals *DTC* and *DTO* can be set for simulation.

In the simulation, vocal tract area functions obtained from an averaging method [5] were used as typical area functions for vowels. Figure 2 shows the vocal tract area functions for five vowels.

Thresholds for glottal rest area A_{g0}

The vocal tract shape affects vocal fold vibration because the vocal tract acts as a load to the glottis. Thresholds for the glottal rest area were determined by the following procedures.

(1) Calculate the vocal tract shape when the area of an acoustic tube having minimum cross-sectional area equals 0.16 cm^2 as a threshold value for the area.

(2) Determine the boundary for the occurrence of vocal fold vibration by controlling the glottal rest area from wide to narrow while the vocal tract is fixed to the shape determined from procedure (1).

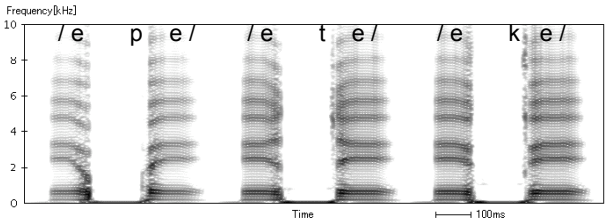


Figure 5: Sound spectrograms for /epe/, /ete/ and /eke/.

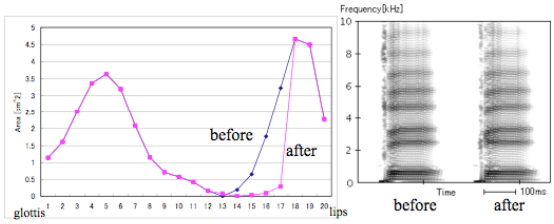


Figure 6: Area functions for /k/ before and after modification and sound spectrograms for /ke/ in /eke/.

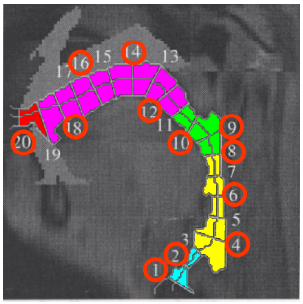


Figure 7: MRI data during the utterance /a/ and the contour of the vocal tract shape in the midsagittal plane.

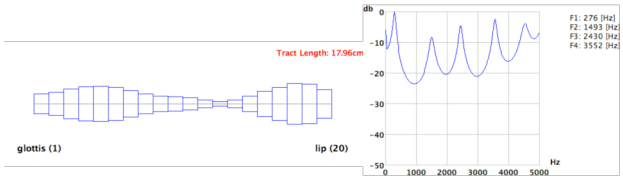
In procedure (2), sound spectrograms and time series data of volume velocity at the glottis were used to determine the boundary.

Figure 3 shows thresholds for glottal rest are A_{g0} for phonetic sequence /eCV/(C= p, t, k, V= i, e, a, o, u). The thresholds vary around 0.14 cm^2 depending on vowels. For simplicity, an average value over the five vowels is used as a threshold for each consonant.

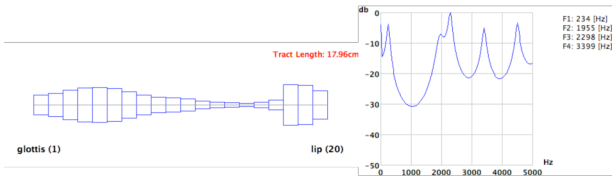
Simulation for relative intervals

Figure 4 shows time ranges of *DTO*s for the phonetic sequence /eCV/ mentioned above. Minimum *DTO* corresponds to the boundary for producing explosion. The range is determined by auditory impression. As shown in the figure, the consonant /p/ has lager minimum *DTO*s than the consonants /t/ and /k/. This tendency is in agreement with the result concerning articulatory movements previously reported [2].

Figure 5 shows sound spectrograms for the synthesized speech /epe/, /ete/ and /eke/. Energy distribution in frequency during explosion is reasonable for /epe/ and /ete/, lower frequency for /p/ and higher one for /t/. However, energy distribution for /k/ has higher frequency components. Synthesis of /eke/ will be discussed in the next section.

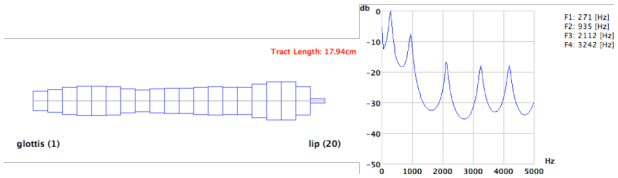


(a) before modification

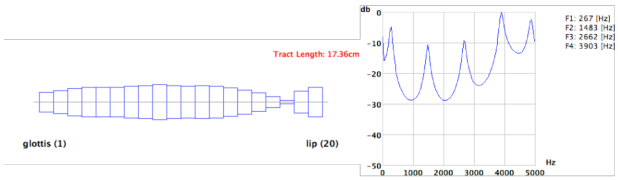


(b) after modification

Figure 8: Similar vocal tract shapes to those of consonant /k/ before and after modification and their transfer functions.



(a) consonant /p/



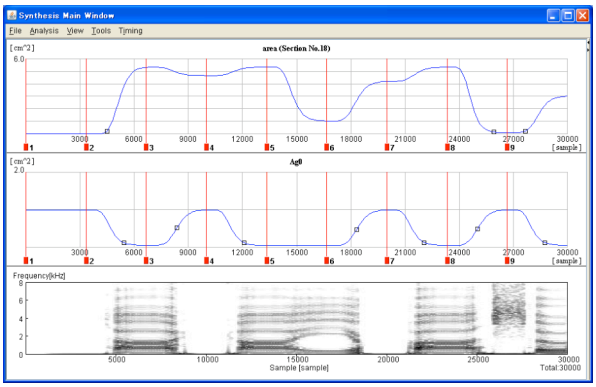
(b) consonant /t/

Figure 9: Similar vocal tract shapes to those of consonants /p/ and /t/ and their transfer functions.

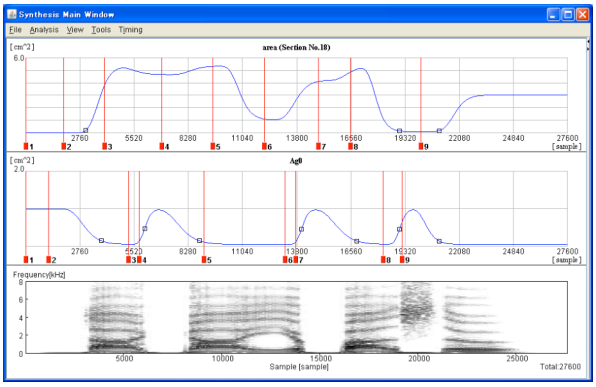
CONSIDERATION OF VOCAL TRACT AREA

As shown in Figure 5, the synthesized consonant /k/ has energy in a higher frequency range. The synthesized speech /eke/ sounds like /ete/, and its quality is low. Therefore, the area function for /k/ was modified. Figure 6 shows area functions for /k/ before and after modification and their sound spectrograms. In the modification, setting areas around the acoustic tube of section 15 smaller results in the increase of energy level around 2 kHz as shown in the spectrograms. Figure 7 shows an example of MRI data during the utterance /a/ and the contour of the vocal tract shape in the midsagittal plane. The acoustic tubes near section 15 in Figure 6 roughly correspond to the point numbered 15 in Figure 7. Although the MRI data in Figure 7 is not for /k/, the figure provides information on the place at which the modification for /k/ in Figure 6 occurs in the vocal tract.

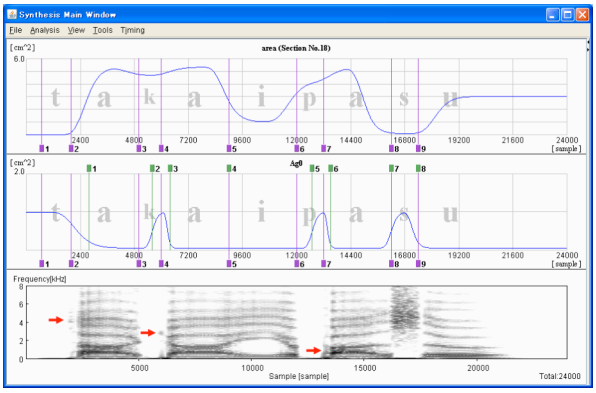
In order to investigate the detail of the phenomenon, a vowel synthesis system [6]-[8] was used to evaluate the effect of the modification on the vocal tract transfer function. Figure 8 shows similar vocal tract shapes to those of consonant /k/ before and after the modification and their transfer functions. Each vocal tract does not have a complete closure at the point of articulation compared with area functions in Figure 6. Comparing with these two transfer functions in Figure 8, we can observe an increase of the energy level caused by the approach of second and third formant frequencies. The modification of the area function provides increase of the second



(a) Initial view.



(b) View window after adjustment.



(c) View window for duration after adjustment.

Figure 10: An example of the adjustment process for /takaipasu/ through the GUI window.

formant frequency and decrease of the third one and results in the increase of energy level around 2 kHz. Because the distribution of energy during stop consonants is a cue of the recognition, the simulation suggests that the modification of the area improves the quality of sound /k/ through the energy distribution. Although considered vocal tract areas used in the simulation with the vowel synthesis system are not identical to those used in Figure 6, the system provides useful insight for resonant frequencies of the vocal tract just after the explosion.

Figure 9 shows similar vocal tract shapes to those of consonants /p/ and /t/ and their transfer functions. The transfer functions provide information on the energy distribution just after the release of the closure of the vocal tract. As shown in Figure 5, we can see energey distribution in lower frequency for /p/ and in higher frequency for /t/.

SYNTHESIS SYSTEM WITH PARAMETER ADJUSTMENT FUNCTION

A GUI system with a parameter adjustment function has been developed based on the simulation results mentioned above. In this section, an adjustment procedure and an example are shown. The outline of the adjustment procedure is as follows:

(1) Input phoneme sequence

(2) Set initial state (Figure 10 (a))

(3) Set duration for each phoneme

(4) Calculate the value of the time constant for each phoneme

(5) Calculate the time at the boundary of adjoining phonemes by evaluating area parameters

(6) Adjust input timing of the command for each phoneme so that the generated time pattern for area follows the boundary conditions mentioned before

(7) Adjust input timing of the hypothetical motor command for parameter A_{g0} to set the relative timing between articulation and vocal fold vibration (Figure 10 (b))

Figure 10 shows an example of the adjustment process through the GUI window; (a) initial view of the GUI window, (b) view window after adjustment, (c) duration view after adjustment. In this example, phonemic sequence /takaipasu/ is synthesized. In the initial view (a), area and A_{g0}, are displayed as functions of time. The vertical lines show hypothetical motor commands that generate the changes of the parameter values as shown in Figure 1 (b). After inputting the values *DTC*s, *DTO*s and durations for the phonemes to text boxes that appear in the setting process, input time of the hypothetical motor commands are justified so that the relative timing between articulation and vocal fold vibration follows the inputted values by the adjustment procedure mentioned above. Changes in the timings of the commands are shown in the area and A_{g0} windows in Figure 10(b).

A view window that indicates duration for each phoneme is also included in the system. In this view mode, a user can confirm the duration for each phoneme as shown in Figure (c). The bottom of each figure shows a sound spectrogram of the synthesized speech calculated from inputted values of the related parameters. We can see differences in sound spectrograms between before and after adjustment, especially for /ta/, /ka/ and /pa/. Durations for vowels and consonants are successfully controlled depending on the parameters set through the GUI.

The automatic adjustment function described here allows us to produce stop consonants /p/, /t/ and /k/.

CONCLUSION

In this study, simulation for investigating relative timing between articulation and vocal fold vibration in the production for stop consonants /p/, /t/ and /k/ was performed with a speech synthesis system using a transmission line model. The simulation results provide useful information on automatic and successful adjustment parameter setting involved in controlling the speech synthesis system. The system is useful not only for speech synthesis but also for making synthesized speech materials for auditory experiments, because it allows us to design synthesized speech through the GUI.

The authors would like to thank Naoki Otsuka for the development of a prototype system. Part of this work was supported by Grant-in-Aid for Scientific Research ((C)20560398) from the Ministry of Education, Science, Sports and Culture, Japan.

REFERENCES

1 K. Ogata and Y. Sonoda, "Development of a GUI-based articulatory speech synthesis system", Proceedings of International Conference on Spoken Language Processing (ICSLP2002), Denver, USA, pp.1517–1520(2002)
2 K. Ogata and Y. Sonoda, "Quantitative analysis of articulatory behavior based on cascaded first-order systems", Acoust. Sci. & Tech. **23**, 117–120(2002)
3 M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-35**, 955–967(1987)
4 K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Tech. J., **51**, 1233–1268(1972)
5 K. Ogata and B. Yang, "A study of vocal tract area functions for speech synthesis system based on transmission line model", Proceedings of 19th International Congress on Acoustics (ICA 2007), Vol.CD-ROM (cas-03-010.pdf) pp.1–6(2007)
6 K. Ogata, "A Web-based articulatory speech synthesis system for distance education", Proceedings of Interspeech 2005, pp.1049–1052(2005)
7 K. Ogata and Y. Shin, "Web education contents for vowel production using vowel synthesis system", Proceedings of 8th International Conference on Information Technology Based Higher Education and Training, Kumamoto, pp.525–529(2007)
8 https://grebe.eecs.kumamoto-u.ac.jp/~http/manual_page/