DEVELOPMENT OF A GUI-BASED ARTICULATORY SPEECH SYNTHESIS SYSTEM

Kohichi Ogata and Yorinobu Sonoda
Faculty of Engineering
Kumamoto University, Kumamoto
ogata@eecs.kumamoto-u.ac.jp

ABSTRACT

The authors have developed a speech synthesis system based on articulation in order to study relationships between the human vocal tract shapes and speech signals. The articulatory speech synthesizer proposed by Sondhi and Schroeter is used to produce speech sounds in the system. By using a GUI (Graphical User Interface) technique, the system can provide easy and interactive operation for users. Therefore, the system is useful not only for speech research but also for teaching materials in speech science. In this paper, effectiveness of the system for understanding articulatory-to-acoustic relationships is shown for steady vowels. Modification of the system for producing continuous vowels is also shown with more realistic description of vocal tract shapes.

1. INTRODUCTION

An articulatory speech synthesis system, which simulates speech production process, has a potential for producing very natural speech output. The realization of such a system requires the understanding of articulatory-to-acoustic relationships. In order to study relationships between the human vocal tract shapes and speech signals, the authors have developed a vocal tract simulator whose transfer function is determined from the shape of a vocal tract [1]. Speech synthesizer used in our system is based on the speech production model proposed by Sondhi and Schroeter [2]. A GUI-based system can be an effective and user-friendly tool for teaching materials in speech science as well as speech research. In this paper, the effectiveness of the developed system is shown for steady and continuous vowels.

2. SYSTEM CONFIGURATION

Figure 1 shows the configuration of the system. Speech synthesis and analysis modules were combined into the GUI-based system by using a script language Tcl/Tk [3]. Tcl/Tk is an easy-to-learn language and is available for many computer platforms; UNIX, MS-Windows, Macintosh, etc. In addition, it allows programmers to expand its function by linking with other programming languages such as C language. By using its linking function, C and FORTRAN programs for numerical calculation in speech synthesis and analysis were integrated into the GUI system. The version of Tcl/Tk used in the system was 8.0. The development of the system was mainly performed on a workstation of HP VISUALIZE C160L.

Speech synthesis and analysis modules were interfaced by inverse Fourier transformation and digital convolution. Therefore, the synthesizer is able to describe a natural acoustic interaction between glottal source and the load provided by the vocal tract. In the synthesizer, two-mass model of the vocal cords proposed by Ishizaka and Flanagan [4] was used to produce the glottal source. The excitation of the glottis is controlled by vocal-cord tension (Q), glottal rest area (Ag0), and subglottal (lung) pressure (Ps). The vocal tract is approximately described by cascaded 20 acoustic tubes and is parameterized by cross-sectional area of each tube and the vocal tract length.

Figure 1 System configuration.

Figure 2 Articulatory speech synthesizer.

(PA7300LC 160MHz) with a HP-UX10.20 operating system. Considering recent progress of personal computers, we are now porting the system to a notebook computer(Pentium III 600MHz) with a Red Hat Linux 7.0.13 operating system. A USB audio interface is used for high quality audio output.

Speech synthesizer used in our system is based on the speech production model proposed by Sondhi and Schroeter [2]. The model consists of three parts: vocal cords, vocal tract and lip models. Figure 2 shows a schematic diagram of the synthesizer. The synthesizer is based on a hybrid method where the glottis is modeled in the time domain and the vocal tract is modeled in the frequency domain. These two models are interfaced by inverse Fourier transformation and digital convolution. Therefore, the synthesizer is able to describe a natural acoustic interaction between glottal source and the load provided by the vocal tract. In the synthesizer, two-mass model of the vocal cords proposed by Ishizaka and Flanagan [4] was used to produce the glottal source. The excitation of the glottis is controlled by vocal-cord tension (Q), glottal rest area (Ag0), and subglottal (lung) pressure (Ps). The vocal tract is approximately described by cascaded 20 acoustic tubes and is parameterized by cross-sectional area of each tube and the vocal tract length.
Determining vocal tract shape

Numerical data from MRI technique

A six-parameter articulatory model proposed by Ishizaka

Determining vocal cord oscillation

Vocal cord tension $Q$ related to pitch frequency
- Constant
- Pitch fluctuation from real speech

Synthetic Speech

Fig. 3 A block diagram of speech synthesis.

Figure 3 shows a block diagram of speech synthesis. In this procedure, a vocal tract shape is determined through a display window and then the glottal parameters are set to produce synthetic speech sounds. The system supports several input methods for determining the vocal tract shape. Recently, vocal tract shapes during the production of steady vowels can be obtained relatively easily by using MRI techniques. Vocal tract parameters extracted from MRI images are available in the system. A six-parameter articulatory model proposed by Ishizaka\[5\], which provides a vocal tract shape, is also utilized for determining the values of cross-sectional areas for 20 sections. In addition, the vocal tract shape displayed in the window can be modified by using the mouse and the keyboard.

3. SYSTEM APPLICATIONS TO ARTICULATORY-TO-ACOUSTIC STUDY

Recent MRI techniques provide us with information on the geometric configuration of the vocal tract during speech. However, it is not easy to extract the exact contours of vocal tract near the glottis or the lips because of the problems of image slice thickness, acquisition time of images, and S/N ratio etc. In such cases, an articulatory simulator like our system is useful for studying the effects of changes in the vocal tract area functions on the resultant acoustic properties of the tract.

Here, an example of system applications to the articulatory-to-acoustic relationships is shown for steady vowel /u/.

Figure 4 shows vocal tract shapes approximated by twenty acoustic tubes and their transfer functions for vowel /u/. In this example, which has a problem in image processing, the transfer function of the initial vocal tract shape (a) based on the MRI data is different from that of real speech, especially the second formant frequency. From a comparison of the figures from (a) to (d), changing the cross-sectional areas near the glottis and/or the lips has good influences on the transfer function of the vocal tract. It is clear that the improvement of the transfer function consists of lowering of a formant frequency F4 by increase of the area near the glottis and lowering of F2 and F3 by decrease of the lip area.

Thus, the system developed here provides us with useful information on articulatory-to-acoustic relationships.
4. SYSTEM IMPROVEMENTS FOR CONTINUOUS VOWELS

Synthesis of continuous vowels is available in the system. In the synthesis, area functions of vocal tracts for steady vowels obtained from MRI data were used and the transition of the vocal tract from a vowel to a successive vowel was described by an interpolation function. A model composed of cascaded first-order systems is used as an interpolation function of vocal tract movements. Since the response time-pattern of the cascaded first-order systems can describe the trajectories of articulatory movements with high accuracy [6], the model was used to describe the transitions.

Figure 5 shows the configuration of the cascaded first-order systems and articulatory movement patterns based on the step responses of the cascade model; displacement, velocity, and acceleration. The impulse response of the model is given by

\[ h_n(t) = \frac{A}{(n-1)!T} \left( \frac{T}{t} \right)^{n-1} e^{-\frac{t}{T}}, \quad t > 0 \]  

(1)

where \( n \) is the number of cascaded components, that is, the order of the model, \( T \) is the time constant, and \( A \) is the amplitude. As shown in Fig.5(b), a step input as a hypothetical motor command to the model produces an ascending motion. Successive ascending and descending motions are obtained by combining two responses as shown in Fig.5(c).

Figure 6 illustrates a change in the area of one of acoustic tubes resulting from articulation. In this case, increase of the area is described based on a step response of the cascaded first-order systems. For simplification, it is assumed that each acoustic tube varies its area perpendicular to the axis depending on the step response of the model. Figure 7 shows an example of the change in the area of the fifth tube from the glottis for continuous vowels /aoiue/. The transition of the area is smoothly described. In the figure, long vertical lines show the time instants at which inputs to the cascaded systems as hypothetical motor commands shown in Fig.5(b). These time instants can be modified by using a computer mouse. Therefore, the system allows us to control speaking rates. Figure 8 shows (a) a change in vocal tract shape and (b) a sound spectrogram for synthesized continuous vowels /aoiue/. Since the function based on cascaded first-order systems can generate smooth change in the area functions as shown in the figure, good synthetic sounds were obtained.
Approximating the vocal tract by cascaded acoustic tubes as shown in Fig.4 has the advantage of reducing the calculation costs in the speech synthesis because of its simple configuration of the tract. However, a more realistic view of the vocal tract shape allows us to understand the geometry of the tract during continuous speech. In order for a more realistic view of the vocal tract, mid-sagittal view of the vocal tract from MRI data was used to form a vocal tract shape as shown in Fig.9. The $\alpha$--$\beta$ model [7] was used to transform the area into the sagittal distance in each section of the vocal tract. Change in the vocal tract shape from the initial /a/ to the middle /i/ in the utterance /aia/ shows a more realistic view of vocal tract configuration. The residual standard deviations for the $\alpha$--$\beta$ model approximation were less than about 0.5 cm$^2$ except for the sections near the glottis and the lips. Although the description of the tract shape at the lips is not enough in the figures, the displayed vocal tract shape shows a typical constriction of the vocal tract for each vowel.

5. CONCLUSIONS

As shown here, the GUI-based speech synthesis system developed here is an effective tool for understanding speech production process. Production model for consonants is studied for further development of the GUI system.

The authors would like to thank Mamoru Furuta for his help with OpenGL programming. Part of this work was supported by Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture, Japan.

6. REFERENCES


Fig.9 Examples of vocal tract shapes for /a/ to /i/ transition during the utterance of continuous vowels /aia/. 