Mandarin vowel synthesis based on 2D and 3D vocal tract model by finite-difference time-domain method

Yuguang Wang*, Hongcui Wang*, Jianguo Wei* and Jianwu Dang *,†
*School of Computer Science and Technology, Tianjin University,
92 Weijin Road, Nankai District, Tianjin 300072, China
E-mail: heureux@126.com
†School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: jdang@jaist.ac.jp

***Abstract*—Finite-difference time-domain (FDTD) method is an effective numerical method to do acoustic simulation. This paper focused on the details of Mandarin vowel synthesis based on 2D and 3D vocal tract model by FDTD method. To do so, a 3D vocal tract shape and vocal tract area function were extracted from the MRI volumetric images during Mandarin vowel production. 3D and 2D model with staggered FDTD mesh were constructed based on the vocal tract and its area function, respectively. Finally, vowels were synthesized by simulating wave sound propagation in the vocal tract using FDTD method with the two-mass vocal folds model. The formant frequencies of synthesized vowels were compared to those of real speech sounds. It is found that the mean absolute errors of formant frequencies were 7.77% and 6.07% for 2D and 3D model, respectively. Results suggested that both 2D and 3D model are capable of producing speech formants in about the same accuracy. However, 3D method exhibits more realistic phenomenon in high frequency region because it was based on complete 3D vocal tract model. It is also observed that the bandwidths of real speech can be achieved through setting the normal sound absorption coefficient within a proper range.**

I. INTRODUCTION

Human speech sound can be replicated to a nearly realistic level using technologies such as linear prediction (LP) [1], corpus-based concatenative synthesis [2]. In concatenative synthesis, phonemes are extracted from the recorded speech corpus and arranged together to construct new words and sentences. These methods only focus on the sound. They ignore the process of the speech sound generation. However, articulatory speech synthesis use computational techniques to synthesize speech based on modeling the human vocal tract and the articulation processes.

Many numerical calculation methods have been used to simulate sound wave propagation in the vocal tract for years, such as transmission line modeling (TLM), finite element modeling (FEM) [3], digital waveguide mesh modeling (DWM) [4] and finite-difference time-domain (FDTD) [5]. Jack (2004) used the 1D and 2D digital waveguide for acoustic simulation, and concluded that 2D mesh exhibits greater realism [4]. Among these methods, FDTD is an accurate and fast method for acoustic characteristics analysis of vocal tract and speech synthesis. Takemoto (2010) performed detailed acoustic analysis of the 3D vocal tract for five Japanese vowels by FDTD method and the results indicated that the FDTD method successfully simulated the acoustic phenomena within a physical model of the vocal tract up to 10 kHz [5]. So far, no further studies have investigated the details of vowel synthesis based on the two or three dimensional FDTD method.

In this paper, we focus on the Mandarin vowel synthesis based on the vocal tract extracted from magnetic resonance imaging (MRI) images by 2D and 3D FDTD method. In section Ⅱ, we briefly introduced the extraction of 3D vocal tract and its area function. Based on the vocal tract, the 3D and 2D staggered FDTD mesh were constructed. In section Ⅲ and Ⅳ, we described the basic principle of FDTD method and the implementation details of FDTD algorithm. In section Ⅴ, the acoustic characteristics of the synthesized vowels were analyzed. Finally, conclusions were drawn in section Ⅵ.

II. EXTRACTION OF 3D VOCAL TRACT AND ITS AREA FUNCTION

The MRI database of Chinese mandarin production aims to provide high resolution vocal tract data for Chinese spoken language research. 3D MRI data during Mandarin vowel production was used in the paper. To obtain a 3D vocal tract shape, the first thing to do is to perform image preprocessing and teeth superimposition on the MRI volumetric images. The images were converted from DICOM to TIFF and denoised using ImageJ software, which is released by NIH (National Institutes of Health, USA). We referred to a method proposed by Takemoto to visualize the teeth in MRI images which is the teeth superimposition approach [6]. Fig. 1 (a) shows the MRI image of the mid-sagittal plane of Chinese vowel /a/ after teeth superimposition. Due to the high quality of the MRI images, we can easily extract vocal tract by threshold segmentation algorithm. A vocal fold line was at the mid-sagittal plane to separate the vocal tract from the sub-glottal tract system. Then, we constructed the volume data of the 3D vocal tract using a region growing algorithm. The size of the voxel of the vocal tract is $0.5 \times 0.5 \times 0.5$ mm^3. Fig. 1 (b) shows the extracted 3D vocal tract shape of /a/ by a male

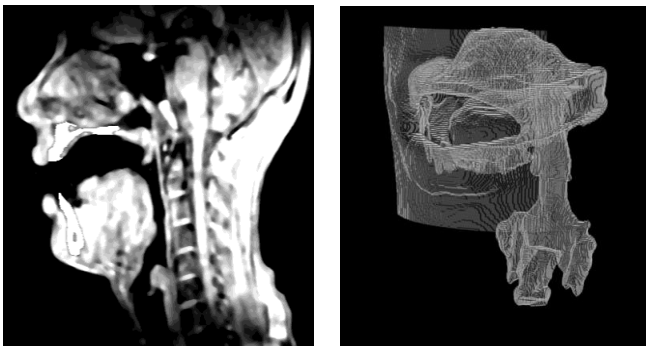

Fig. 1 (a) The MRI image of the mid-sagittal plane during phonation of /a/; (b) 3D Vocal tract of /a/ extracted from MRI images.

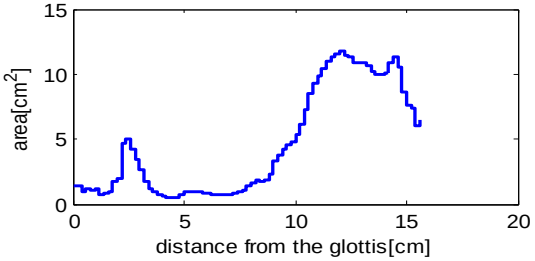

Fig. 2 The area function of the vocal tract of /a/.

subject. The nasal cavities are excluded in the vocal tract in this paper. The 3D vocal tract contains the main tract tube and side branches, which includes the piriform fossae, epiglottic valleculae, and inter-dental spaces.

The vocal tract area function, which shows the relationship of the cross-sectional area along with the distance from the glottis to the mouth, describes the length and the shape of the vocal tract. After the extraction of the vocal tract, area function can be easily calculated from the vocal tract. Firstly, a central line of the vocal tract was drawn on the mid-sagittal plane [7]. Then, 3D vocal tract shape was sliced as a series of cross-sectional areas perpendicular to the line. The area function was obtained by measuring each of the cross sectional area. Fig. 2 shows the area function of Chinese vowel /a/.

III. THE FINITE-DIFFERENCE TIME-DOMAIN METHOD

FDTD is a numerical method proposed by Yee [8] to solve the electromagnetic wave propagation problem. Recently, it has been frequently used in simulating the wave propagation in acoustics [5, 9]. In order to solve the absorption problem of the outgoing waves near the boundary of the analysis field, perfectly matched layers (PML) [10] are placed around it. Since the formulas of 3D FDTD method are quite similar to 2D FDTD method, we only introduce the basic equations next.

The three-dimensional wave equations considering the energy loss are given as follows:

$$\nabla p(x,y,z,t) = -\rho \frac{\partial}{\partial t} u(x,y,z,t) - \alpha^* u(x,y,z,t) \tag{1}$$

$$\nabla \cdot u(x,y,z,t) = -\kappa \frac{\partial}{\partial t} p(x,y,z,t) - \alpha p(x,y,z,t) \tag{2}$$

where p is the pressure field and u is the velocity of particle, κ ($\kappa=1/\rho c^2$) is the compressibility of the medium, ρ is the mass density of the medium, c is sound velocity in the medium, α is the usual compressibility attenuation in acoustic medium. The attenuation coefficient associated with density α^* ($\alpha^*=\alpha\rho/\kappa$) is generally zero in the acoustic medium.

If we denote all functions of discrete space and time as

$$f(i\Delta x, j\Delta y, k\Delta z, n\Delta t) = f^n(i,j,k) \tag{3}$$

then wave equations can be formulated as Equation (4) and (5) using the central difference principle.

$$p_x^{n+\frac{1}{2}}(i,j,k) = \frac{2\kappa - \alpha\Delta t}{2\kappa + \alpha\Delta t} p_x^{n-\frac{1}{2}}(i,j,k) - \frac{2\Delta t}{(2\kappa+\alpha\Delta t)\Delta x}(u_x^n(i+\frac{1}{2},j,k) - u_x^n(i-\frac{1}{2},j,k)) \tag{4}$$

$$u_x^{n+1}(i+\frac{1}{2},j,k) = \frac{2\rho - \alpha^*\Delta t}{2\rho + \alpha^*\Delta t} u_x^n(i+\frac{1}{2},j,k) - \frac{2\Delta t}{(2\rho+\alpha^*\Delta t)\Delta x}(p^{n+\frac{1}{2}}(i+1,j,k) - p^{n+\frac{1}{2}}(i,j,k)) \tag{5}$$

$$p^{n+\frac{1}{2}}(i,j,k) = p_x^{n+\frac{1}{2}}(i,j,k) + p_y^{n+\frac{1}{2}}(i,j,k) + p_z^{n+\frac{1}{2}}(i,j,k) \tag{6}$$

where the pressure p is a scalar variable which is the sum of the pressure along x, y and z direction. The equations for u_y, u_z, p_y, p_z are omitted here, since they have the same shape as u_x and p_x.

To simulate the wave reflection on the vocal tract wall, we use a simple method proposed by Yokata [9]. The particle velocity on the boundary can be formulated as follows:

$$u_x(i+1/2,j,k) = \frac{p^{n+1/2}(i,j,k)}{Z_n} n_x \tag{7}$$

$$u_y(i,j+1/2,k) = \frac{p^{n+1/2}(i,j,k)}{Z_n} n_y \tag{8}$$

$$u_z(i,j,k+1/2) = \frac{p^{n+1/2}(i,j,k)}{Z_n} n_z \tag{9}$$

$$Z_n = \rho c \frac{1+\sqrt{1+\alpha_n}}{1-\sqrt{1+\alpha_n}} \tag{10}$$

where (n_x, n_y, n_z) is the normal vector of the boundary point near the vocal tract wall, α_n is the normal sound absorption coefficient and Z_n is the normal acoustic impedance on the boundary.

IV. IMPLEMENTATION DETAILS

A. Generating vocal tract mesh

Both the voxel data of 3D vocal tract model and the FDTD method use orthogonal system, which is a big advantage for FDTD simulation. We can directly generate the Cartesian mesh for the vocal tract and place PML layers around the analysis field. For the 2D case, Cartesian mesh was

constructed based on the area function. Once the spatial intervals in x and y direction are set, the grid number based on the length of the vocal tract and the area function can be calculated. The width of each cross section at the 2D mesh is set to be the radius of the corresponding cross section of the vocal tract.

B. Experiment parameters

The only difference between the 2D and 3D FDTD simulation is the spatial and temporal interval. For the 2D simulation, we chose $\Delta t=4\times10^{-6}$s as the temporal interval and $\Delta x=\Delta y=0.002$m as the spatial interval. For the 3D case, we chose $\Delta t=2\times10^{-6}$s as the temporal interval and $\Delta x=\Delta y=\Delta z=0.002$m as the spatial interval.

The other experiment setups for both cases are the same. 12 layers of PML medium were placed around the field to absorb the outgoing waves. Sound speed and medium density for the simulation are listed as follows:

$$C_{air}=346.3m/s,\rho_{air}=1.17kg/m^3$$

$$C_{wall}=1500.0m/s,\rho_{wall}=1000.0kg/m^3$$

The vocal tract can be assumed as time-invariant system, and transfer function of it can be achieved by setting up a Gaussian pulse, as shown in Equation (11), at the glottis.

$$p(t)=\exp\{-((n\cdot\Delta t-T)/0.29T)^2\},t=n\cdot\Delta t \quad (11)$$

where $T=0.646/f_0$, $f_0=5$kHz, and n is the time step.

To generate a speech sound, we used the voice source instead of the Gaussian pulse. In this experiment, we used the two-mass model to generate the voice source. Voice source was set up at the glottis. An observation point was chosen near the mouth to record the waveform. FFT with 65536 points was used to calculate the transfer function from the observed waveform. All the simulations were done on a workstation with Windows system, which has four 3.09 GHz cores and 4 G memory.

V. RESULTS AND ANALYSIS

In the experiment, we use Gaussian pulse as input source to generate impulse response of the vocal tract and adopt two-mass model glottal wave to synthesize vowels. Real speech of the corresponding vowels is recorded in a sound proof room. The first four formant frequencies of vowel /a/ extracted from the real speech are 681Hz, 1101Hz, 2794Hz, and 3803 Hz. Formant frequencies synthesized by FDTD method were compared to the real speech.

A. Vowel simulation

Figure 3 shows the speech waveform synthesized by 2D FDTD method and the impulse response (also referred as transfer function) of the vocal tract for Chinese vowel /a/. F1-F4 of the impulse response, which are labeled in the figure, are 621Hz, 1231Hz, 2644Hz, 3609Hz.The error rates of first formant frequencies compared to the real speech are -8.8%, 12.0%, -5.4 and -5.1%. Mean absolute error is 7.77%.

Figure 4 shows the waveform synthesized by 3D FDTD method and the transfer function of the vowel /a/. The first four formant frequencies are 679Hz, 1076Hz, 2626Hz,

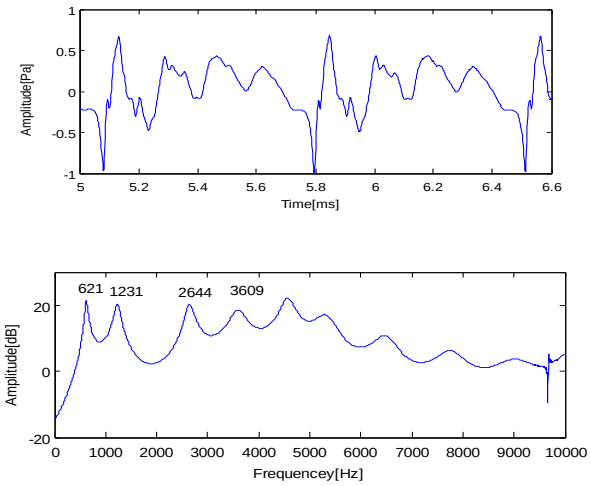

Fig. 3 (a) Speech waveform synthesized by 2D FDTD method. (b) Transfer function of Chinese vowel /a/ by 2D FDTD method.

3206Hz, respectively. The corresponding error rate of the formant frequencies compared to the real speech is -0.3%,-2.3 %, -6.0, -15.7%. Mean absolute error is 6.07%.

The speech waveform and the formant frequencies simulated by 2D and 3D method are quite similar to real speech. The absolute mean error is smaller than Wang's (2008) study, which is about 10% [11]. If you listen to the vowels synthesized by two methods, you'll find the one by 3D is more real than that by 2D model. That's because the 3D simulation exhibits more detailed information. In the 3D FDTD transfer function, the dip around 4 kHz caused by the piriform fossa is very clear [12], while there are no dips around 4 kHz in the 2D case. The big disadvantage of 3D simulation is a high calculation cost. It took about 160 minutes to synthesize a period of 0.2s speech, while the 2D simulation only needs 31 seconds for the same one.

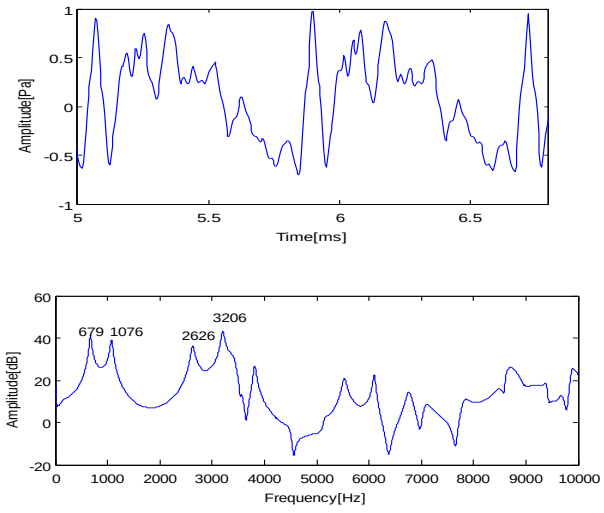

Fig. 4 (a) Speech waveform synthesized by 3D FDTD method. (b) Transfer function of Chinese vowel /a/ by 3D FDTD method.

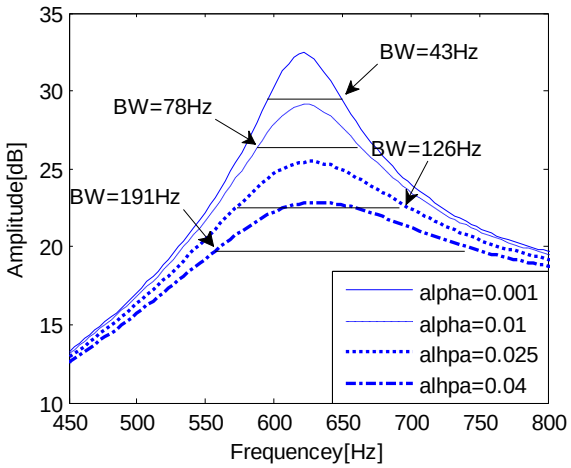

Fig. 5 Bandwidth variations of the vowel /a/ by 2D FDTD. "BW" is short for bandwidth. "Alpha" is the normal sound absorption coefficient.

B. Formant bandwidths

The formant of the speech includes three factors: the formant frequency, bandwidth and amplitude. The most important factor is the formant frequency, which has been discussed in the previous section. In this section, we are going to discuss the second important factor crucial to the reality of the speech: the bandwidth of the formant. As for the FDTD method, the formant bandwidths are directly influenced by the normal sound absorption coefficient. Fig. 5 shows the first formant bandwidth variations of the vowel /a/ by 2D FDTD method. As one can see in the figure, formant bandwidths are quite sensitive to the absorption coefficient. As the absorption coefficient of the wall rises from 0.001 to 0.04, the bandwidth increases from 43Hz to 191Hz. Therefore, it is proper to set the coefficient between 0.01 and 0.025 to reach the general bandwidth of real speech, which is about 80-120Hz. The simulation method of the formant bandwidths for 3D FDTD method is similar to the 2D case.

VI. CONCLUSIONS

In this study, we briefly described the extraction of 3D vocal tract and area function from MRI data and investigated the details of Chinese vowel synthesis by 2D and 3D model using FDTD method. The synthesized vowels were compared to the real speech. Results suggest that both methods can produce accurate speech formants. As we know, the simulation by 2D model is much faster than that by 3D, sometimes it is more suitable for synthesize speech using 2D FDTD method in practice. But, the 3D simulation uses complete 3D vocal tract, it is better to do detailed and accurate acoustic analysis of vocal tract. At last, the control of formant bandwidth was also investigated. It is observed that formant bandwidths are quite sensitive to the absorption coefficient.

REFERENCES

[1] Schnell, K., Lacroix, A., "Time-varying linear prediction for speech analysis and synthesis," ICASSP, pp. 3941-3944, 2008.

[2] T. Dutoit, "High-quality text-to-speech synthesis: an overview," J. Elect. Electron. Eng. Australia: Special Issue on Speech Recognition and Synthesis, vol. 17, no. 1, pp. 25–37, 1998.

[3] K. Aoyama, H. Matsuzaki, N. Miki and Y. Ogawa, "Finite element method analysis of a three-dimensional vocal tract model with branches," *J. Acoust. Soc. Am.*, 100, 2657–2658 (1996).

[4] J. Mullen, D. H. Howard and D. T. Murphy, "Acoustical Simulations of the Human Vocal Tract Using the 1D and 2D Digital Waveguide Software Model," The 7th International Conference on Digital Audio Effects (DAFx-04), Naples, Italy, 2004.

[5] H. Takemoto, P. Mokhtari, T. Kitamura, "Acoustic analysis of the vocal tract by finite difference time domain method," *Journal of the Acoustical Society of America*, vol. 128, Issue 6, (2010).

[6] Takemoto, H., Kitamura, T., Nishimoto, H., and Honda, K. (2004). "A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions," Acoust. Sci. & Tech. 28, 33–38.

[7] Takemoto, H., K. Honda, et al. (2006). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data." Journal of the Acoustical Society of America 119(2): 1037-1049.

[8] K. S. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas Propag*. AP-14, pp. 302–307, (1966).

[9] T. Yokota, S. Sakamoto and H. Tachibana, "Visualization of sound propagation and scattering in rooms," *Acoustical Science and Technology*, Vol. 23, No. 1, pp.40-46, (2002).

[10] J. P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves," *J. Comput. Phys*. 114, 185–200, (1994).

[11] Gaowu Wang, Tatsuya Kitamura, Xugang Lu, Jianwu Dang and Jiangping Kong (2008). "MRI-based Study of Morphological and Acoustical Properties of Mandarin Sustained Steady Vowels", Journal of Signal Processing, Vol.12, No.4, pp. 311-314, July 2008

[12] J. Dang, K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," Journal of *Acoustical Society of America*, vol. 101, pp. 456–465, (1997).